2012

Using topographic and soils data to understand and predict field scale soil moisture patterns

Zachary James Van Arkel
Iowa State University

Using topographic and soils data to understand and predict field scale soil moisture patterns

by

Zachary James Van Arkel

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Agricultural Engineering

Program of Study Committee:
Amy Kaleita, Major Professor
Brian Hornbuckle
Lie Tang

Iowa State University

Ames, Iowa

2012

TABLE OF CONTENTS

LIST OF EQUATIONS

ABSTRACT

Identifying and understanding the impact of within-field soil moisture patterns is currently limited by the time and resources required to do sufficient monitoring. The spatial and temporal variance of soil moisture complicates the ability to monitor and effectively predict soil moisture values. Remote sensing offers non-invasive techniques to measure soil moisture, but the resolution is too coarse to be of immediate value in many of the applications requiring soil moisture information. Obtaining high resolution soil moisture data requires dense sensor networks to adequately monitor changing spatial and temporal soil moisture patterns. The aim of this study is to develop methods to estimate soil moisture values at the field scale without the need for exhaustive pre-sampling. This is achieved by finding critical sampling locations within the field based upon topographic and soils data that can adequately predict field scale soil moisture. Given these sampling locations and values for soil moisture at those points, an interpolation method is developed that is independent of the spatial relationship between the sampling locations and the points to be interpolated. Ultimately, these approaches can be used as a method to find critical sampling points and interpolate field-scale soil moisture values based upon topographic and soils data that can be collected in a one pass operation and thus eliminate the need for extensive soil moisture monitoring.

CHAPTER 1: GENERAL INTRODUCTION

Introduction

Soil moisture (θ) is a key component in weather prediction, crop growth simulation, and environmental performance modeling. Compared to other sinks in the hydrologic cycle, the volume of soil moisture is small (~0.001% of global water), but it is of fundamental importance to many hydrological, biological and biogeochemical processes (USGS 2012). The surface soil water content is important because it controls the energy exchange between the atmosphere and land surface. Knowing that soil moisture is an important variable in understanding terrestrial hydrology, obtaining soil moisture measurements has become a focus for researchers in environmental modeling.

The current methods for measuring θ are reviewed in Robison et al. (2008). From this research, the measurement of θ can be generalized into two different methods: remote sensing and *in-situ* θ sensors. By definition, remote sensing makes measurements to θ without being in direct contact with the soil. Remote sensing technologies have the ability to measure θ at a variety of different spatial and temporal scales. Satellite remote sensing devices can cover large areas in a short amount of time, but have low spatial resolution (~15-40 km pixel size). Airborne remote sensing methods provide smaller pixel size (~10m), but operation is expensive and measurements are weather permitting. In-field remote sensing technologies can provide smaller resolutions, but measurements are limited to the field in which the instrument is installed.

Apart from remote sensing, ground-based sensors provide θ values at the point scale. Because only values for a single point are provided, networks with a high number of sensors

are required to understand spatial θ patterns. Besides taking time to install these sensor networks, the money required to purchase and maintain the sensors make the method inefficient for larger scale spatial θ estimations.

A common problem encountered in sensing θ is the lack of spatial resolution and timely values required to monitor the highly variable values of θ. Remote sensing techniques can cover large areas, but the spatial resolution is inadequate for many of the models requiring θ information. Sensor networks provide high resolution data, but are expensive to maintain and the values of θ are only valid for the area in which those sensors are installed. With these challenges in mind, the research objectives of this thesis are to:

1. Identify optimal soil moisture sampling locations based upon readily available field data that can then be used to estimate field-scale θ values with the same accuracy as a sensor network.
2. Estimate θ at the sub-field scale depending on the relationship of the topographic and soils data between the optimal θ sampling points identified in objective 1, and the unknown points within the field to be interpolated.

Ultimately, achieving objectives 1 & 2 will eliminate the need for dense sensor networks to find the spatial patterns of θ at the sub-field scale and thus save time and money needed to purchase, install, and monitor a large number of ground based θ sensors. Furthermore, achieving these objectives would help in bridging the gap between the different scales at which θ readings are available from remote sensing and *in-situ* measurements. Ideally, researchers will be able to accurately estimate θ values at scales needed without the

need for dense *in-situ* sampling networks. Knowing that sensor networks are time and resource inefficient, the challenge then becomes determining how to accurately estimate θ values at the same resolution as a sensor network with a fewer number of sampling locations. Finding the number of samples that need to be taken to adequately estimate θ values at the scale desired, and then deciding where to locate the sampling stations is the topic of Chapter 2. Methods for deciding the number of sampling locations and in finding the location for sampling stations are introduced and tested on fields where sensor networks have been installed and monitored.

The next challenge in bridging the gap between the different scales of θ measurement techniques is how to estimate θ values at unknown points in the landscape given the optimal sampling locations found in Chapter 2. Different landscape characteristics allow θ values to change abruptly making interpolation of θ values difficult. A new method is needed for interpolation of θ values that is dependent of the spatial relationship between the point with known θ and the point to be interpolated. A new method is proposed in Chapter 3 that relies solely on topographic and electromagnetic inductance data of the soil to interpolate θ values at unknown points within the landscape. From this method, θ values are closely related to the different topographic and physical indices having significant impact on θ that are introduced in the literature review.

Literature Review

Because soil moisture (θ) is an important variable, much research has been devoted to finding the factors that have an effect on θ patterns in an effort to find estimation methods. The complexity and variety of landscapes used in the current research leads to differing and sometimes conflicting results. Finding those factors most influential on spatial θ patterns is the key to understanding and modeling soil moisture. The factors having an impact at the scale to which this study is concerned are discussed in this review.

Estimation of θ patterns would not be complete without the inclusion of topographic features. The topography of the landscape has an impact on flow channels, infiltration, potential radiation, and is related to the different soil types. Numerous studies include different topographic characteristics in attempts to model and predict θ (Yoo and Kim 2004; Western et al. 2001; Wilson et al. 2004; Famiglietti et al. 1998; Kim and Barros 2002; Mohanty and Skaggs 2001). Though each of these studies were completed on different spatial scales, all use the influence of topographic features in θ estimation. Famiglietti et al. (1998) provides an in depth analysis of different topographic indices, how they are computed, and why they have an impact on θ patterns.

The movement of water due to gravitational potential is the basis for the influence of topography on soil moisture. Studies have shown that θ data is inversely proportional to the elevation. (Henninger et al. 1976; Robinson and Dean 1993; Crave and Gascuel-Odoux 1997). Weeks and Wilson (2006) note that it is typical to find higher moisture contents near the toe of the slope than at the crest. At the field scale it is possible to have higher elevations within the field that exhibit higher θ values. This is may be due to high elevations that are flat

and collect water or a variety of other different factors. Such factors can complicate the inverse relationship between θ and elevation thus requiring other topographic information to adequately estimate soil moisture.

Elevation data can be used to compute a variety of different topographic indices. The slope at a point is a function of the elevation at the point in question and the elevation of the surrounding points. Studies have shown the influence of slope values on soil moisture variability (Hills and Reynolds 1969, Moore et al. 1988). The slope value is important because it determines the drainage characteristics, the amount of infiltration, and thus the runoff produced. A steep slope discourages infiltration whereas a low slope value encourages infiltration or evaporation from that point. As with all topographic indices, the scale at which the slope is found is important to identify. Because the slope can be measured over a variety of different lengths, clarifying the scale at which the slope is calculated helps in understanding its impact on θ variability.

The curvature value is another index that has an influence on θ values. In general, the curvature is the measure of concavity or convexity of the landscape (Famiglietti et al. 1998). A correlation between curvature and soil moisture was documented in Moore et al. (1988). Tomer et al. (2006) found that surface curvature was the terrain attribute most commonly correlated with soil moisture. Concave landscapes will pool water because they have upslope contributing area. In contrast, points in a convex landscape have a smaller upslope contributing area. A convex shape will shed water resulting in lower soil moisture values. A landscape lacking curvature (a plane) will likely shed water in similar ways over the entire area. Understanding the curvature at a point is important in understanding the organization of soil moisture values at that point.

Researchers identify wetness indices that are a function of topography to help in prediction of θ patterns. Western et al. (1999) found a wetness index that is a function of the upslope contributing area to be the best univariate predictor of θ patterns at times when the mean moisture content of the field was high. Beven and Kirby (1979) introduce the steady state wetness index which is a function of the upslope contributing area and the slope of the point. Similar to the results from Western et al. (1999), this index was more successful in explaining θ values at times when the field had a high mean moisture content.

Soil properties

The different hydraulic properties of different soil types will have an impact on θ patterns. The hydraulic properties of the soil are closely related to the soil texture and structure. Brady and Weil (1999) cite the importance of soil texture saying it 'clearly exerts a major influence on soil moisture retention.' Other studies confirm that soil texture has a significant impact on soil moisture content (Hawley et al. 1983, Henninger et al. 1976, Crave and Gascuel-Odoux 1997). The large particles of sand result in a smaller surface area for the attachment of water molecules. In contrast, the small size of clay particles provides a large surface area for attachment. The different sizes of particles within the soil affect the ability of water to infiltrate and percolate, and affects the ability of water to be evaporated and transpired by plants. In addition to the soil texture, the soil structure also has an impact on the soil moisture variability. A well aerated, well granulated soil has more pore space and therefore greater holding capacity for water. Compact soils will have smaller pores that limit infiltration and hold water for longer periods of time. Soil structure is also related to the

existence of macroporosity which is a controlling influence of moisture movement within the soil (Niemann and Edgell 1993).

Electromagnetic inductance (EMI) data is gaining popularity in precision agriculture applications for identifying changes in soil type. The electrical conductivity correlates strongly with the soil particle size and texture and thus can be tied to θ patterns (Tromp and McDonnell 2009; Grisso et al. 2009). The connection between soil texture and particle size with the hydraulic characteristics make EMI a valuable index in predicting θ. Khakural et al. (1998) found a linear relationship between electrical conductivity and soil water profile storage. Huth and Poulton (2007) found that EMI can provide quick and efficient means for monitoring θ in agroforestry systems. The connection between θ and the EMI data will be used in this research.

Potential radiation

A factor affecting θ that takes both the topographic information and the soil properties into account is the potential radiation. In *Soil Physics*, Horton and Jury note that the rate of evaporation from a wet, bare soil surface is a function of external meteorological conditions including wind speed, relative humidity, and the flux of radiant energy (2004). The flux of radiant energy is a function of the soil albedo and slope aspect at a point. Horton and Jury then go on to introduce different stages of evaporative loss from the soil. In the initial stage when the soil surface is wet, the evaporative loss occurs at the maximum rate, after drying, the evaporation is then controlled by other factors determined by the soil.

The slope aspect and the albedo of the soil are the key factors in determing potential radiation at a point in the landscape. The slope aspect is a function of the elevation and is

determined by the direction of the slope. The direction of the slope influences the solar irradiance thus influences potential radiation (Famiglietti et al 1998). The albedo of the soil also has an influence on potential radiation because of its impact on the amount of energy received from the sun. The differing color of soils will affect the amount of radiation absorbed at the surface. A dark loam soil will soak up more radiation from the sun and thus have higher evaporation rates than that of a light clay soil. During much of the growing season when the θ is monitored, the soil is covered by a crop canopy, thus eliminating the effect of albedo on θ. Nevertheless, θ will likely be influenced by the albedo of the soil during the early segments in the growing season and after crops have been harvested and the soil surface layer is exposed.

Western et al. (1999) noted that potential radiation was the best predictor of soil moisture during dry periods. Jackson et al. (1967) noted the effect of slope, aspect, and albedo on potential evaporation from hillslopes. Reid (1973) found a correlation between the aspect and soil moisture. Weeks and Wilson (2006) point out that north-facing slopes in the northern hemispheres receive significantly less radiation than horizontal or south facing slopes and the authors develop a method to predict the soil radiation at a point and thus predict potential evaporation. The potential radiation and its effect on evaporation may be one of the most important factors in determining surface θ values.

Vegetation

Because almost two-thirds of the water falling on the earth's surface is returned to the atmosphere via transpiration, it is not a surprise that the vegetation cover has an influence on θ values (Brady and Weil 1999). The control and effect of vegetation on θ changes

depending on the vegetation type, density, and season (Famiglietti et al. 1998). Lull and Reinhart (1955) found that θ variability increased with decreasing canopy coverage. Besides transpiring the water in the soil, vegetation changes the pattern at which moisture falls on the surface of the soil. In forests the majority of the water hitting the canopy flows down the trunks of trees. Similarly, in row crop fields with a canopy, water runs down the stem of the plant and thus has an impact on spatial patterns of θ (Brady and Weil 1999). Different vegetation types provide different amounts shade and change the pattern of airflow over the soil. This impacts the potential for evaporation from the soil surface under the vegetation. Because this research is concerned with θ patterns at the field scale, it is assumed that the vegetation is homogenous over the study areas and thus will not be a determining factor in θ variability. Nonetheless, it is important to note the impact of vegetation on within field patterns.

Mean moisture content

The mean soil moisture content of the field also has an impact on θ patterns. Henninger et al. (1976) and Hawley et al. (1982) found the variance of θ decreased with decreasing mean moisture content. Hills and Reynolds (1969) argued that θ variability would be highest in the middle range of mean moisture content. In the middle range, moist areas could be present at the same time as dry areas. Whereas after a rainfall, all areas would be saturated, decreasing the variability of θ. In contrast, Western et al. (1999) found that θ patterns exhibit a high degree of organization during wet periods (high mean soil moisture) and a low degree of organization during dry periods. Famiglietti et al. (1998) found that variability in θ decreases with decreasing mean θ content. Yoo and Kim (2004) found that

the influence of soil properties and topographic features increases after rainfall. In a study to generate spatial patterns of θ, Wilson et al. (2005) based their methods on static topographic features and changed the model depending on different wetness conditions. Chang (2001) notes that it is necessary to connect the interdependencies of soil properties, topography, and mean soil moisture content when attempting to predict θ values. The inconsistency of the results of these studies make it difficult to identify when the mean moisture content will be the most influential in identifying θ patterns.

Combined influences

The most pressing difficulty apparent is the variety of different studies identifying different factors having the most influence on θ patterns. Famiglietti et al. (1998) found that during wet conditions, soil moisture is most strongly characterized by porosity and hydraulic conductivity of the soil (both of which are soil properties). During dry conditions, a correlation with soil moisture is more controlled by elevation, aspect, and clay content. Western et al. (1999) found that during wet conditions soil moisture was most influenced by topography. Kaleita et al. (2007) found no conclusive relationships between overall θ patterns and topographic and soil indices. Different findings from previous research can be explained to an extent by differences in climate, soils, vegetation, topography, and the sampling time period. Even if a strong correlation can be found between terrain indices and θ, Western et al. (1999) point out that a significant amount of random behavior exists within the θ continuum which cannot be predicted.

As can be inferred from the above information, θ is not a factor of only one topographic, physical, or chemical characteristic. Finding a univariate predictor would be

beneficial for interpolation methods, but the variability in the soil hydrology system leads to a variety of factors that have an influence. Similar to other complex natural processes a combination of factors influence the variability of soil moisture at different scales. Kaleita et al. (2007) found that stable spatial patterns of soil moisture are linked to a combination of topography, particle size, and drainage pattern. Herbst et al. (2006) were best able to predict soil hydraulic properties at a point given the relative elevation, the slope, and the slope aspect. Although not attempting to predict θ, Green et al. (2007) used elevation, slope, aspect, curvature, and upslope contributing area in combination with spatial coordinates to predict crop yield. Mohanty and Skaggs (2001) noted the need to develop quantitative relationships between θ and various soil, topographic, and vegetation characteristics. Wilson et al. (2005) found a variety of terrain indices that had predictive power of θ patterns. In their concluding remarks, the authors state that spatial distribution of θ is not based on one terrain index but on a weighted combination of indices. Similarly, Western et al. (1999) describe an "index approach" where a variety of different indices are found for points throughout the landscape and used to determine θ behavior. A combination of indices is needed to accurately estimate the dynamic behavior of θ. In the end, the difficulty of modeling natural processes provides uncertainty and increases complexity. Given the main factors affecting the θ variability, the goal of this research is to employ the most dominate physical parameters having an effect on spatial variability of soil moisture to estimate θ values.

Self-organizing maps

Given the complex behavior of soil moisture and the variety of factors having an influence on its value, a method of analysis is needed that can effectively evaluate a data set with many variables. One of the methods that will be used in this research is a type of artificial neural network called a self-organizing map. Self-organizing maps (SOMs) were first developed in the early 1980s by Teuvo Kohenen. The goal behind producing the algorithm was to map similar patterns (pattern vectors close to each other in the input signal space) onto contiguous locations in the output space (Kohonen 1995). Early algorithms were used in speech pattern recognition, but since their inception have been applied to numerous data sets in many fields of research. Unlike other classification techniques, SOMs do not require the class of the input vector to be known. This allows the user to input data with unknown classes into the algorithm and then identify classes based upon the output map.

Input vectors are presented to the SOM and the vectors then 'self-organize.' The output of the SOM algorithm is a two dimensional map made up of 'neurons.' Input vectors are assigned to neurons and are then displayed on the output map to show the relationship between different input vectors. Based upon the the values of the variables for each of the input vectors within the neuron, a vector for each neuron within the input space can be calculated. This process is described in detail below.

A set of p input observation vectors, $\boldsymbol{x}_{input} = [x_1, x_2, \dots, x_n] \in \Re^n$ is fed into the SOM. Input vectors are compared to a set of N neurons on the output layer, $\boldsymbol{m}_i = [m_{i1}, m_{i2}, \dots, m_{iN}] \in \Re^N, (i = 1,2, \dots, N)$. Each input pattern is compared to each output neuron on the output 2-dimensional map. The winning neuron (the neuron to which the input vector is assigned) is chosen based on the formula:

$$\|\boldsymbol{x}_{\boldsymbol{input}} - \boldsymbol{m}_c\| = min_i\{\|\boldsymbol{x}_{\boldsymbol{input}} - \boldsymbol{m}_i\|\} \quad (1.1)$$

where $\boldsymbol{m}_c$ is the winning neuron. This represents the minimum Euclidean distance between the input vector and 'winning' neuron on the output map to which the input vector is assigned. After finding the winning output neuron for each input vector, the neurons on the map are updated according to the following equation:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{ci}(t)[\boldsymbol{x}(t) - \boldsymbol{m}_i(t)] \quad (1.2)$$

where t denotes the index of the iteration step, $\boldsymbol{x}(t)$ is the vector input sample of $\boldsymbol{x}_{input}$ in the iteration t, and $h_{ci}(t)$ is called the neighborhood function around the winning neuron c. During training, $h_{ci}(t)$ is a decreasing function of the distance between the *ith* and the *cth* model on the map node. After the presentation of each input vector, the region around the best matching vector (determined by $h_{ci}(t)$) to is stretched towards $\boldsymbol{x}(t)$. For convergence it is necessary that $h_{ci}(t)$ goes to 0 when t goes to ∞. The end result is that neighboring neurons on the output grid have similar weight vectors in the input space.

The number of neurons in the output map and the dimensions of the map can be chosen arbitrarily or can be determined by the number of input vectors. Vesanto et al. (2000) suggest the number of neurons should be $5\sqrt{n}$ where n is the number of input vectors. Given this rule for finding the number of neurons to be used in the output map, the dimensions of the map (height and width) then must be chosen to correspond to that number of neurons. One method of determining the size of the map is by finding the two largest eigenvalues of the training data. After finding the ratio between those two values the ratio between the length and width of the map is set to that ratio. The actual length and width is adjusted so that their product is similar to the number of map units determined by the rule above.

Another method for determining the size of the map is to minimize quantization and topographic error. Quantization error is the average distance between each data vector and its 'best matching unit' or neuron vector. Topographic error quantifies the number of data vectors for which the best matching unit is not adjacent (Cereghino and Park 2009). In this method of sizing, different maps of different sizes are constructed and the map with minimum values for quantization and topographic error is chosen.

The SOM algorithm can be changed depending the desires of the user. The map units and the size of the map can be manipulated, different distance formulas can be used to find the winning neuron, and different neighborhood functions can be chosen to change the region that is 'stretched' towards the input vector. Changing all of the above factors in the algorithm will change the resulting output map and how it is organized. Finding the optimal map for the specific application is the challenge of the researcher. More details about the SOM algorithm can be found in Kohonen (1995).

K-means clustering

In order to partition the multivariate structure of the neurons in a SOM and in the input data, K- means clustering will be utilized. MacQueen (1967) first introduced the K-means clustering algorithm as a tool to classify and analyze multivariate observations. In the K-means algorithm, the initial 'means' of a decided upon number of clusters (k) is randomly selected from the input data set. Clusters are created by associating each input vector to the nearest mean with the following formula:

$$z_n = min_{i \in 1 \ldots k}[d(x_n, m_i)] \tag{1.3}$$

where z_n is the cluster to which the input vector x_n is assigned, $d(\ )$ is the distance calculated between the input vector and the different k means. Similar to the SOM algorithm, different distance algorithms can be used for finding the difference between vectors. The vectors are then partitioned and the geometric center of each of the clusters becomes the new mean. The geometric center is found with the following formula:

$$m_k = \frac{1}{N_k} \sum_{n:z_n=k} x_n \tag{1.4}$$

where m_k is the mean of the k^{th} cluster, N_k is the number of points assigned to the k^{th} cluster, and x_n is the input vector. After the new means are found in (4), (3) is calculated and this process continues until assignments $z_{1:N}$ do not change. The random selection of an input vector for the initial means of the clusters has an impact on the resulting cluster assignments of the input vectors. Due to this result, it is important to run the algorithm multiple times to validate the resulting clustering assignments.

The K-means clustering algorithm will be used in this research to partition both the SOM neuron data and the input data. This is done to evaluate the value of the SOM algorithm in assigning the input data to neurons with common characteristics that are subsequently used to cluster the data. Besides partitioning the data into neurons, the SOM algorithm allows a visual interpretation of the input data that is qualitatively valuable in describing the relationship between the different input data. Both the SOM algorithm combined with the K-means algorithm and the K-means algorithm alone are valuable in their ability to handle multivariate data as is used in this research.

References

Brady, N. C., Weil, R. R., The Nature and Properties of Soils, 12th ed., 1999.

Beven, K. J., and N. J. Kirkby, A physically based variable contributing area model of basin hydrology, Hydrological Science, 24, 43-69, 1969.

Cereghino, R., and Y. S. Park, 2009, Review of the Self-Organizing Map (SOM) approach in water resources: Commentary, Environmental Modeling & Software, 24, 2009.

Chang, D. –H., Analysis and modeling of space-time organization of remotely soil moisture (Ph.D. dissertation), University of Cincinnati, 2001.

Crave, A., Gascuel-Odoux, C., The influence of topography on time and space distribution of soil surface water content, Hydrological Processes, 11, 203-210, 1997.

Famiglietti, J. S., Rudnicki, J. W., Rodell, M., Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas, Journal of Hydrology, 210, 259-281, 1998.

Green, T. R., J. D. Salas, A. Martinez, and R. H. Erskine, Relating crop yield to topographic attributes using spatial analysis neural networks and regression, *Geoderma*, 139, 23-37, 2007.

Grisso, R., M. Alley, D. Holshouser, and W. Thomason, Precision Farming Tools: Soil Electrical Conductivity, Virginia Cooperative Extension Publication 442-508, 2009.

Hawley, M. E., Jackson, T. J. McCuen, R. H., Surface soil moisture variation on small agricultural watersheds, Journal of Hydrology, 62, 179-200, 1983.

Henninger, D.L., Peterson, G.W., Engman, E. T., Surface soil moisture within a watershed: Variations, factors influencing, and relationships to surface runoff, Soil Science Society of American Journal, 40, 773-776, 1976.

Hills, T. C., Reynolds S. G., 1969. Illustrations of soil moisture variability in selected areas and plots of different sizes, Journal of Hydrology, 8, 27-47.

Horton, R., Jury, W. A., Soil Physics, 6th ed., 2004.

Huth, N. I., and P. L. Poulton, An electromagnetic induction method for monitoring variation in soil moisture in agroforestry systems, *Soil Research*, 45, 63-72, 2007.

Jackson, R. J., 1967. The effect of slope, aspect and albedo on potential evaporationfrom hillslopes and catchments, Journal of Hydrology, 6, 60-69.

Kaleita, A. L., Hirschi, M. C., Tian, L. F., Field-Scale Surface Soil Moisture Patterns and Their Relationships to Topographic Indices, Transactions of the ASABE, 50, 2007.

Khakural, B. R., P. C. Robert, and D. R. Hugins, Use of non-contacting electromagnetic inductive method for estimating soil moisture across a landscape, *Communications in Soil Science and Plant Analysis*, 29, 1998.

Kim, G. and A. Barros, Downscaling of remotely sensed soil moisture with a modified fractal interpolation method using contraction mapping and ancillary data, *Remote Sensing of Environment*, 83, 400-413, 2002.

Kohonen, T. *Self-Organizing Maps*. Springer, Verlag Berlin Heidelberg New York, 1995.

Mohanty, B. P., and T. H. Skaggs, Spatio-temporal evolution and time-stable characteristics of soil moisture within remote sensing footprints with varying soil, slope, and vegetation, *Advances in Water Resources*, 24, 1051-1067, 2001.

Lull, H.W., Reinhart, K.G., Soil moisture measurement. U.S.D.A. Southern For. Exp. Sta., New Orleans, LA., Occas. Paper No. 140, 1955.

Moore, I. D., Burch, G. J., Mackenzie, D. H., Topographic effects on the distribution of surface water and the location of ephemeral gullies, Trans. Am. Soc. Agric. Eng., 31, 1098-1107, 1988.

Niemann, K. O., and M. C. R. Edgell, Preliminary analysis of spatial and temporal distribution of soil moisture on a deforested slope, Physical Geography, 14, 449-464, 1993.

Reid, I., The influence of slope orientation upon the soil moisture regime, and its hydrogeomorphical significance, *Journal of Hydrology*, 19, 309-321, 1973.

Robinson, D., C. Campbell, J. Hopmans, B. Hornbuckle, S. Jones, R. Knight, et al., Soil Moisture Measurement for Ecological and Hydrological Watershed-Scale Observatories: A Review, *Vadose Zone Journal,* 7(1), 358-389, 2008.

Robinson, M. Dean, T.J., Measurement of near surface soil water content using a capacitance probe, Hydrological Processes, 7, 77-86, 1993.

Tomer, M. D., Cambardella, C. A., James, D. E., Moorman, T. B., Surface-Soil Properties and Water Contents across Two Watersheds with Contrasting Tillage Histories, Soil Science Society American Journal, 70, 620-630, 2006.

Tromp-van Meerveld, H. J., and J. J. McDonnell, Assessment of multi-frequency electromagnetic induction for determining soil moisture patterns at the hillslope scale, *Journal of Hydrology*, 368, 56-67, 2009.

U.S. Geological Survey, Where is Earth's water located?, *Water Science for Schools*, available online <http://ga.water.usgs.gov/edu/earthwherewater.html>, 2012.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, SOM Toolbox for Matlab 5, Technical Report A57, Neural Networks Research Centre, Helsinki University of Technology, Helsinki, Finland, 2000.

Weeks, B. and G. W. Wilson, Prediction of evaporation from soil slopes, *Canadian Geotechnical Journal*, 43, 2006.

Western A., G. Bloschl and R. Grayson, Toward capturing hydrologically significant connectivity in spatial patterns, *Water Resources Research*, 37, 83-97, 2001.

Western, A. W., Grayson, R. B., Bloschl, G., Willgoose, G. R., McMahon, T. A., Observed spatial organization of soil moisture and its relation to terrain indices, Water Resources Research, 35, 797-810, 1999.

Wilson D., A. Western, and R. Grayson, Identifying and quantifying sources of variability in termporal and spatial soi moisture observations, *Water Resources Research*, 40, 1-10, 2004.

Wilson D., A. Western, and R. Grayson, A terrain and data-based method for generating the spatial distribution of soil moisture, *Advances in Water Resources*, 28, 43-54, 2005.

Yoo, C. and S. Kim, EOF analysis of surface soil moisture field variability, *Advances in Water Resources*, 27, 831-842, 2004.

CHAPTER 2: FIELD-SCALE SOIL MOISTURE ESTIMATION USING SELF ORGANIZING MAPS AND K-MEANS CLUSTERING TO IDENTIFY CRITICAL SAMPLING POINTS

A paper to be submitted to *IEEE Transactions on Geoscience and Remote Sensing*

Zach Van Arkel and Amy Kaleita

Abstract

Identifying and understanding the impact of field-scale soil moisture patterns is currently limited by the time and resources required to do sufficient monitoring. This study uses self-organizing maps (SOMs) and K-means clustering algorithms to find critical sampling points to estimate field-scale soil moisture. Points within the field are clustered based upon topographic and soils data and the points representing the center of those clusters are identified as the critical sampling points. Using soil moisture information from the critical sampling points and the number of points within each cluster, a weighted average is found and used as the estimate mean field-scale soil moisture. Field-scale soil moisture estimations from this new method are compared to the techniques introduced by Vachaud et al. (1985) to find optimal sampling locations based upon temporal soil moisture data. Ultimately, the new approach can be used to find critical sampling points to estimate soil moisture measurements without the need for exhaustive pre-sampling.

Introduction

The modeling of hydrologic processes is a key component in weather forecasting, crop growth simulation, and environmental performance prediction. Compared to other sinks

in the hydrologic cycle, the volume of soil moisture (θ) is small, but it is of fundamental importance to many hydrological, biological and biogeochemical processes. Knowing that θ is an important variable in these processes, having access to accurate θ information is of value to researchers in environmental modeling.

Current techniques for measuring θ are presented in a review by Robinson et al. (2008). Finding an efficient method for measurement at the resolution required is the challenge for applications where θ is an important input. Techniques vary from *in-situ* sensing instrumentation at the smallest spatial and temporal scales, to remote sensing satellites that provide θ information over large areas with less frequency. With each different method for measurement comes variance in the cost of the sensor, the cost of installation, the amount of maintenance required, the accuracy of the sensor, the ease of use, and the depth at which the θ is measured.

On a global scale, the most efficient technique for gathering θ data is using remote sensing. The constant motion of the satellite allows large areas to be covered with frequencies adequate for weather and crop models needing the θ information. The launch of the SMOS (Soil Moisture Ocean Salinity) satellite and the upcoming launch of the SMAP (Soil Moisture Active Passive) satellite will efficiently produce large amounts of θ data. With the idea that θ information from satellites is going to be readily available, the need to validate the accuracy of the satellite readings without extensive pre-sampling of θ arises. Knowing that θ readings from the satellite correspond to readings from the ground-based sensors is important in maintaining consistency and accuracy in the modeling applications requiring θ data. Because a large sensor network is required to check the accuracy of θ measurements from the satellite resolution, identifying representative sampling points throughout the

landscape that adequately estimate θ at the satellite resolution is one of the keys to validating remotely sensed data.

Current methods for field-scale estimation require extensive time-series θ measurements from a network of *in-situ* sensors. One of the most common methods for finding optimal sampling locations to estimate θ at the field scale is the Rank Stability Analysis (RSA) developed by Vachaud et al. (1985). Given extensive time series θ data, sampling points within the field are identified as optimal sampling locations based upon having the smallest standard deviation of mean relative θ. These points are determined rank stable because they have the smallest variance with respect to the field mean θ. Though this has proven to be a valuable and relatively accurate method for θ estimation, weaknesses of the method make it unattractive. Besides the time and monetary resources required to find the temporal θ data for analysis, the reliance on empirical data is a downfall of the method. Because the method is based solely on empirical data, the ability to recognize why certain locations are better to sample than others is limited to the sampling points used to find the rank stable locations. Additionally, Yang (2010) argued that choosing random points from the sampling grid within the field was as reliable in field-scale θ estimation as the RSA method.

The desire to identify critical sampling points without time-series θ information leads to the need for a new method of data analysis. A method of analysis gaining popularity in modeling natural processes is that of computational intelligence, specifically self-organizing feature maps (SOMs). Similar to the human brain but different than other classification techniques, SOMs learn patterns from complex data sets and then classify information accordingly. Typically, SOM networks learn to cluster groups of similar input patterns from

a high dimensional input space onto a low dimensional lattice of 'neurons' in an output layer (Kohonen 2001). The end result is an output layer (map) with contiguous neurons having similar input patterns (Kalteh et al. 2008). Often, SOMs are combined with a clustering algorithm to find similarly behaving clusters within the input data. The complexity and variation of temporal and spatial θ behavior, and the variety of factors having an impact on θ patterns, promote the use of computational intelligence methods in modeling θ behavior. Because soil moisture patterns are constantly changing over space and time, the ability of SOMs to analyze, cluster, and model data make it an attractive tool for complex natural processes. Complex data sets can be represented in a two dimensional map where similarities between sampling points can be observed. Additionally, because SOMs are unsupervised in nature, the exact class to which a specific sampling point belongs is not required. Data can be organized and clustered without knowing to which class each sampling point belongs. This is attractive because the high variability in θ values in spatial and temporal data sets complicates classification with traditional methods.

Recent research supports the use of SOMs in modeling temporally and spatially varying natural processes. Annas et al. (2007) employed SOMs to identify temporally variable 'hotspots' in an attempt to predict fire risk. Mele and Crowley (2008) applied this method to classify soils based on biological, chemical, and physical properties. Honda and Konishi (2001) incorporated SOMs in their research to cluster cloud images from time-series satellite weather images. Also in a study showing the useful application of SOMs in a temporally varying environment, Lauzon et al. (2004) analyzed θ profiles on temporal scales with wavelet analysis and SOMs. Other studies attempted to model rainfall and runoff rates,

but few studies have addressed the need to understand spatial and temporal θ characteristics using computational intelligence methods.

Knowing that SOMs can handle large amounts of data from a variety of different variables, the factors impacting spatio-temporal θ patterns can be used as inputs into the algorithm. Understanding the topographic and soil physical properties that have an effect on θ at different scales is crucial to understanding spatial θ patterns. Many studies confirm that both soil physical properties and topography control variations of θ over large areas (Chang 2001, Romano and Palladino 2002). Other studies suggest that topography, soil physical characteristics, vegetation, and the climate are key factors that influence θ variations at the watershed scale (Famiglietti et al. 1998; Yeh and Eltahir, 1998; Western et al. 1999). Qiu et al. (2001) reported that on a smaller scale (field) land-use and soil type have a more pronounced control on θ than topography. Western et al. (1999) observed that patterns in θ result from a combination of both surface and subsurface pathways, while in the summer the potential radiation showed the strongest relationship with θ. Famiglietti et al. (1998) found that the dominant influence on θ changes from differences in soil heterogeneity to joint control by topographic and soil properties as the hill slope dried following rain events. Kaleita et al. (2007) found that stable spatial patterns of θ are linked to the combination of topography, particle size, and drainage pattern. Finding the most influential factors in estimating the θ is key to understanding θ patterns.

The complex combination of factors influencing the spatial variability of θ and the ability of SOMs to represent large data sets in a two dimensional map makes this application appropriate for the use of this method. This research uses SOMs combined with K-means clustering, and the K-means clustering algorithm to find critical sampling points for field-

scale θ estimation based on topographic and soil physical data. Finding a link between the physical data of the landscape and the θ behavior will allow field-scale θ to be estimated for validation without the need for extensive on ground θ monitoring. Recent improvements in LiDAR (Light Detection and Ranging) technologies allow accurate high resolution topographic data to be produced. Given this accurate topographic information, different derivatives of topography can be calculated and used as controlling factors of θ that are input in the SOM and K-means algorithm to find critical sampling locations.

The ultimate goal of this research is to develop, with easily attainable data, a practical plan for designating critical θ sampling points within agricultural fields that can accurately estimate the field-scale θ and eventually help in bridging the gap between point measurements and remotely sensed θ data. First, given past time-series θ information, critical sampling points will be found using SOMs with K-means clustering algorithms and used to find a field-scale θ estimation. The estimates will be compared to estimates found using optimal sampling locations identified by the RSA method. Second, the SOM and K-means clustering algorithms will be used to find critical sampling points depending only on topographic and soil physical data as inputs. Assessing the accuracy of the estimates from the SOM and K-means clustering algorithms compared to the RSA method will determine the feasibility of using these new methods for field-scale θ estimation.

Methods

Field data

This study analyzed *in-situ* θ measurements from the Brooks research field in Story County, Iowa. Soil moisture measurement values were taken in a 300 x 250 meter grid (~18

acres) on the field during the growing seasons (summers) of 2004-2008. The spacing between each sampling point is 50 meters and the coordinates of the grid are given in Universal Transverse Mercator (UTM, a mapping projection that gives location in meters from a datum). The elevation in the field varies by approximately 5 meters and the grid covers six different soil types and a variety of different landscape positions throughout the field (Fig. 1). Six points on the north end of the grid were not sampled in 2006 and thus 2006 will not be used for data analysis. The readings were taken with an average interval of approximately 3 days. The daily sampling time period was limited to a maximum of two hours in order to reduce the θ differences due to drying.

The soil moisture value used for analysis is an average of 3 samples taken within a ~0.5 m radius of each sampling location at a depth of 0-6 cm with a ThetaProbe moisture meter (Delta-T Devices, Cambridge UK, marketed in the United States by Dynamax, Inc., Houston, Texas). Values from the probe were then converted to estimates of volumetric θ using a calibration developed for soils on the Des Moines lobe provided by Kaleita et al. (2005). A field calibration based on ThetaProbe measurements combined with gravimetric sampling resulted in a regression coefficient R^2 of 0.77. The θ values given are in *cm^3 (water)/cm^3 (soil-water-air volume).*

In each season data collection with the ThetaProbe began after planting and samples were taken roughly twice a week in the absence of rain. In total there were 99 measurement days for the 2004-2008 growing seasons (less 2006 growing season measurements). As reference for the temporal θ behavior, precipitation data for each of these growing seasons was obtained from the Ames 8 WSW Station (UTM (Zone 15): 435912E, 4652376N; 42.0208 Lat, -93.7741 Lon) from the National Oceanic and Atmospheric Administration

website. Fig. 2 shows the average θ of all grid points with standard deviation shown by error bars combined with the precipitation during the sampling time period.

To calculate topographic indices, elevation data for the Brooks field was obtained using a GPS receiver mounted on an all-terrain vehicle. The vehicle traveled in the north-south direction with approximately 20 m between each pass. Using the elevation data, slope, planar curvature, and slope aspect were derived for each point on the θ sampling grid using Surfer® (Golden Software, Inc., Golden, Colorado). A 10-meter grid of elevation data was generated and then Surfer® routines were used to find the indices. A 10-meter grid was used based upon the finding by Yang (2010) that this scale was adequate to describe field-scale θ patterns. The grid cell containing each of the sampling points was identified and the topographic indices for the sampling points were extracted from this information.

Strobl et al. (2006) explain the indices and their impact on hydrologic patterns in the landscape. The slope is the rate of change in elevation and controls the energy available to propel surface flow. Curvature is a measure of topographic divergence and convergence and thus has an influence on the concentration of water at the surface. Positive values of planar curvature indicate convergent flow whereas negative values of planar curvature indicate divergent flow. The slope aspect indicates the direction of the slope from a point to its surroundings. This value has an influence on direction of flow and also on the potential radiation received at a point. The radiation received is important in determining θ because it impacts evaporation and transpiration.

The known influence of soil types on hydraulic properties calls for the inclusion of a variable that can capture changes in soil texture. In the absence of high resolution soils data, the electromagnetic inductance (EMI) is used as a proxy index to identify changing soil

properties. Both horizontal (H-H) and vertical (V-V) conductances in units of milliSiemens/meter were gathered using an EMI sled pulled by an all-terrain vehicle. EMI data was interpolated with inverse distance weighting for each of the θ sampling locations in the grid based upon the ~20 m resolution data found with the EMI sled.

Rank stability analysis

The methods introduced by Vachaud et al. (1985) are employed to compare the identification and prediction of sampling points from the methods proposed in this paper. Using time-series θ data, the Rank Stability Analysis method finds the relative difference and standard deviation from the grid mean for each grid point. Points with small standard deviation are determined temporally rank stable. Temporally rank stable points are then used as optimal sampling locations because their θ behavior varies the least in time. Time-series data from the 2004 season was used to find 3 sampling points from the grid with the smallest standard deviation of mean relative difference to the field average. These points are deemed the optimal sampling locations (OSLs) for grid average validation because they have consistent behavior over time with respect to the field average θ content. Given the sampling locations with the smallest standard deviation of the mean relative difference, the field mean θ soil moisture is found with the following equation:

$$\bar{\theta}_{est} = \frac{\theta_{OSL}}{1+ \bar{\delta}_{OSL}/100} \tag{2.1}$$

Where $\theta_{OSL}$is the measured volumetric soil moisture content from an OSL on a given day, $\bar{\delta}_{OSL}$ is the mean relative difference from the OSL, and $\bar{\theta}_{est}$ is the estimated mean soil moisture from this OSL on the given day. When more than one OSL is used to determine the

estimated mean field moisture, equal weights are given to the estimated soil moisture value from each point.

To compare the different techniques of identifying critical sampling locations with the different algorithms, temporal θ data from 2004 and topographic and EMI data were used to construct three matrices: $\mathbf{M}_\theta$, $\mathbf{M}_T$, and $\mathbf{M}_E$. Each matrix contained the points in the sampling grid as rows and the rows then represent the input vectors into the algorithms. In the columns of the matrix $\mathbf{M}_\theta$ contained the 2004 θ sampling days as columns. This is the same data used to find optimal sampling locations based upon RSA. Only the temporal θ data from 2004 was included because this method would be similar to the rank stability method of identifying optimal sampling locations based upon a temporal data series and then using the selected sampling locations for future estimation of θ values. The columns of $\mathbf{M}_T$ contained elevation, slope, slope aspect, and planar curvature, and the columns of $\mathbf{M}_E$ contained elevation, slope, slope aspect, planar curvature, H-H EMI, and V-V EMI. Thus, $\mathbf{M}_\theta$ is a 42X24 matrix of θ values was created (corresponding to 24 sampling days from the 2004 season), $\mathbf{M}_T$ is a 42X4 matrix, and $\mathbf{M}_E$ is a 42X6 matrix.

To address the differing scales of the variables in $\mathbf{M}_T$ and $\mathbf{M}_E$, the values in the matrix were normalized before input into the SOM algorithm and K-means algorithm. A linear transformation for normalization was used with the mean value of each variable set to zero. Each variable was normalized with the function below:

$$x' = \frac{(x-\bar{x})}{\sigma_x} \tag{2.2}$$

Where x' is the normalized value at a point, x is the actual value of the variable at the point, $\bar{x}$ is the mean of all the values for the specific variable, and σ_x is the standard deviation of all the values for the specific variable.

Self-organizing maps

Self-organizing maps are useful in their ability to find patterns in complex data sets with a variety of variables. Input vectors (temporal θ or topographic and EMI data from each point) are presented to the SOM and the vectors then 'self-organize.' Each input pattern is compared to an output neuron on the output 2-dimensional map. A schematic diagram of a SOM is given in Fig. 3. A set of *n* observation vectors $\boldsymbol{x}_{input} = [x_1, x_2, \dots, x_n] \in \Re^n$ is fed into the SOM. Neurons on the output map are represented by the vector $\boldsymbol{m}_i = [m_{i1}, m_{i2}, \dots, m_{in}] \in \Re^n, (i = 1,2, \dots, N)$, where *N* is the number of neurons on the output map. When constructing an SOM, the algorithm compares each input vector to each neuron and the winning neuron, $\boldsymbol{m}_c$, is chosen based on the formula

$$\left\|\boldsymbol{x}_{\boldsymbol{input}} - \boldsymbol{m}_c\right\| = min_i\{\left\|\boldsymbol{x}_{\boldsymbol{input}} - \boldsymbol{m}_i\right\|\}. \tag{2.3}$$

After finding the winning output node for each input vector, the neurons on the map are updated according to the following equation:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{ci}(t)[\boldsymbol{x}(t) - \boldsymbol{m}_i(t)] \tag{2.4}$$

where *t* denotes the index of the iteration step, $\boldsymbol{x}(t)$ is the input sample of $\boldsymbol{x}_{input}$ in the iteration *t*, and $h_{ci}(t)$ is called the neighborhood function around the winning neuron $\boldsymbol{m}_c$. During training, $h_{ci}(t)$ is a decreasing function of the distance between the *ith* and the *cth* neuron. For convergence it is necessary that $h_{ci}(t)$ goes to 0 when *t* goes to infinity. Different distance formulas and neighborhood functions can be chosen within the SOM algorithm. For

simplicity in this study, the default parameters within the MATLAB SOM Toolbox 2.0 (Vesanto et al. 2000) were used. More details about the SOM algorithm can be found in Kohonen (1995).

K-means clustering

A more commonly used algorithm than SOM, K-means clustering is used to separate both the SOM map neurons and the temporal and physical data matrices into three different clusters containing points with similar characteristics. In the K-means algorithm, the initial 'means' of a decided upon number of *k* clusters is randomly selected from the input data set. Clusters are created by associating each input vector (grid point data) to the nearest mean. The vectors are then partitioned and the geometric center of each of the clusters becomes the new mean. This process continues until the points converge. The MATLAB SOM Toolbox contains a K-means clustering algorithm that was used for partitioning the neurons in the output layer of the SOM. Similarly, this function was used to partition the grid point data without inputting the data into the SOM. This bypasses the SOM algorithm altogether to find clusters of points with similar characteristics. Readers are referred to MacQueen (1967) for further explanation of the K-means algorithm.

Using the SOM Toolbox, a unified distance matrix (u-matrix) was created using $\mathbf{M}_{\theta}$, $\mathbf{M}_{T}$, and $\mathbf{M}_{E}$. The u-matrix shows the distance between the hexagonal neurons and can be used to identify patterns within the data. By noticing the color difference in the map and using the color scale, one can see the difference in distances between nodes within the U-matrix. Colors corresponding to small numerical values show that the nodes are closely

related, whereas colors corresponding to large numerical values show divisions within the input data.

After applying the K-means clustering algorithm to both the SOM neuron data and then to $\mathbf{M}_\theta$, $\mathbf{M}_T$, and $\mathbf{M}_E$, the centroid vector of each cluster in each method can be found. As the name suggests, the centroids represent the centers of the clusters created. Using the Euclidean distance formula, the input vector with the smallest distance from the centroid can be found. This input vector (grid point) is then deemed the best matching unit (BMU) to the cluster centroid. These BMUs were then used as the critical sampling locations identified by their respective algorithms. The point number of the BMUs to the cluster centroids for each method can be found in Table 1.

To find the estimated average of the field θ using the sampling points identified by the clustering algorithms, a weighted average was found using the BMUs from each method and the number of points in the corresponding cluster. The weighted average can be found with the following formula:

$$\bar{\theta}_j = \frac{\sum_{i=1}^{n} \theta_{BMU_{ij}} * \#\ of\ sampling\ points\ in\ C_i}{total\ sampling\ points\ in\ grid} \tag{2.5}$$

Where $\bar{\theta}_j$ is the estimated mean θ on the *j*th day, $\theta_{BMU_{ij}}$ is the θ value for the BMU to cluster *i* centroid on the *j*th day, C_i is the *i*th cluster, and $i \in \{1,2,3,4\}$.

To compare the accuracies of the estimated field average from the different methods, the average bias, root mean squared error, and the Nash-Sutcliffe efficiency index were calculated. The Nash-Sutcliffe index provides a number from 1 to -∞ with 1 being a perfectly predicting model (Nash and Sutcliffe 1970, McCuen etal. 2006). A value of zero for this

index indicates that the model predictions are as accurate as the mean of the observed data. The index is calculated with the following formula:

$$NS = \frac{\sum_{t=1}^{T}\left(D_o^t - D_e^t\right)^2}{\sum_{t=1}^{T}\left(D_o^t - \overline{D_o}\right)^2} \quad (2.6)$$

Where D_o is the observed value of θ and D_e is the estimated value of θ at time *t*.

Results and Discussion

SOM visual interpretation

The u-matrix in Fig. 4a gives insight into the divisions within the 2004 temporal θ data ($\mathbf{M}_\theta$). The colors given in the color scale correspond to the Euclidean distance between the different hexagonal neurons of the SOM. A division can be seen between the lower third and upper two thirds of the map in Fig. 4a. This division identifies the existence of a cluster with θ behavior that differs from the points located in the upper two thirds of the map. In the top two thirds of the map, the lack of colors corresponding to large Euclidean distance values shows those nodes have similar temporal θ behavior. The lack of a definitive division between the nodes in the top two thirds of the map supports the use of only two different clusters from the 2004 temporal θ data.

The sampling point identification numbers corresponding to the points in Fig.1 are displayed on the map and thus the landscape positions for each of the points within a cluster can be interpreted. As an example, sampling points 17 and 63 are assigned to the same neuron on the SOM. Viewing Fig. 1, this result agrees with the landscape position of these points as they are both in an area where water converges as determined by surrounding elevation values. Similarly, sampling points 9 and 81 are in the same neuron on the SOM and

are both located at the top of a ridge. Going further, the spatial relationship (opposite corners) on the SOM between the neuron containing sampling points 17 and 63 and the neuron containing points 9 and 81 agrees with the idea that these two points should have very different θ behavior. When looking at the landscape, it could be assumed that sampling points 17 and 63 would exhibit higher than average θ values whereas points 9 and 81 would exhibit lower than average θ values. Such a result explains the maximum spatial distance on the output map of the SOM between these two nodes and gives insight into the relationship between other points on the map as compared to these two nodes. Map neurons between these corners likely will contain sampling points with less extreme behavior.

Fig. 4b and 4c exhibits the u-matrices after inputting $\mathbf{M}_T$ and $\mathbf{M}_E$, respectively. In Fig. 4b, colors corresponding to a larger distance are located in the upper left hand corner of the map. This result suggests that the points located in this section of the map have topographic characteristics that are dissimilar to points located in the lower portion of the map. Interestingly, similar to Fig. 4a, Points 9 and 81 are assigned to the same neuron in the U-matrix. Although in Fig. 4a the points are located in the map neuron that is the farthest distance away from surrounding neurons as denoted by the color scale. In Fig. 4a, points 3. 13, and 65 are all in the same neuron that is the farthest distance from any of the surrounding neurons. In Fig. 4b, these same points are scattered throughout the bottom half of the amp showing an inconsistency partitioning the data based only upon topography.

Fig. 4c includes EMI data with the topographic data to construct a u-matrix. Similar to Fig. 4a, a division is denoted by the color scale between the upper two thirds of map and the bottom third of the map. Again, point 9 and 81 are assigned to the same neuron in the corner of the map. Points 3, 13 and 65 are at the bottom of the map farthest away from points

9 and 81 showing more consistency between Fig. 4a and Fig. 4c. Similarly, points 17, 19, 51, and 63 are all in close spatial proximity in Fig. 4a and in Fig. 4c. The correspondence between the location of the points in the output maps constructed from $\mathbf{M}_\theta$ and $\mathbf{M}_E$ lend support to the inclusion of EMI data in estimating soil moisture behavior.

Besides giving insight into the relationship between sampling points, the u-matrices are valuable in identifying divisions within the input data. The resulting u-matrix in Fig. 4a suggests dividing the data into two different clusters because of the division seen between the lower third and upper two thirds of the map. Fig. 4c also supports a division into two clusters. Fig. 4b is more difficult to analyze because of the lack of a definitive division. Colors corresponding to small distance (blue) at the bottom of the map suggest a cluster from that region. Colors corresponding to higher distances start at the upper left corner and continue down towards the lower right corner. This pattern suggests a division between the upper right and left corners. Thus, the matrix in Fig. 4b suggests a division into three clusters of data.

With the overall goal of accurately estimating the field mean θ while eliminating the need for exhaustive pre-sampling, the question then becomes how many points need to be sampled. Ideally, one point within the field could be identified as the critical sampling point and could be used for field scale estimation. The differing landscape and soil characteristics make identifying one point for accurate estimation difficult. In preliminary studies, three critical points were used for field mean estimation based on having a wet, medium, and dry cluster, based on the existence of three predominant soil textures (sand, silt, and clay), and based on the having three common landscape locations (hilltop, sideslope, toeslope). This decision was consistent with research by Chang (2001) who used three classes to estimate soil texture from remote sensing brightness temperature. Knowing that the u-matrices from

$\mathbf{M}_\theta$ and $\mathbf{M}_E$ support a division into two clusters, and knowing that the u-matrix constructed from $\mathbf{M}_T$ suggests a division of the data into three clusters, it was decided to group the data into one, two, three, and four clusters for identification of critical sampling points.

Estimation from 2004 time-series θ data (M_θ)

Table 1 gives the average bias, root mean squared error, and Nash-Sutcliffe coefficient of efficiency indices from the methods with $\mathbf{M}_\theta$ as input data. When using the RSA method, the RMSE values and NSCE values support the use of a higher number of points in estimating the field mean θ value. Estimation from critical sampling points identified by the SOM K-means and K-means algorithms have lower RMSE values and higher NSCE values when 2 points are used for estimation instead of 3 points. The most accurate estimations are from the 4 critical sampling points identified by both the SOM K-means and K-means algorithms.

The resulting statistical indices support the use of the SOM K-means and K-means methods for identifying critical sampling points over the RSA method. Average bias, RMSE and NSCE values from two critical sampling points identified by the SOM K-means and K-means algorithm show a more accurate estimation than from the 4 point estimation using the RSA method. Average bias and RMSE values are lower for estimation from the 3 points identified for sampling by the SOM K-means and K-means algorithms in comparison to estimation using the 4 RSA OSLs. The improvement in the statistical indices presented support the use of these new methods for finding critical sampling locations from temporally varying θ data.

Estimation from physical data (M_T & M_E)

Given the θ estimation results from the three different methods using $\mathbf{M}_\theta$ and the resulting sampling locations advised by those methods, it is important to realize that all of the above methods used 2004 temporal θ from the Brooks field to identify sampling points. The objective of this research was to find an efficient and accurate method for finding θ at the field scale without the need for exhaustive pre-sampling of θ as would be required by the above methods. The success that was found with the SOM K-means and K-means methods with the θ data gave confidence in applying those methods to find sampling locations based on topographic and physical soil data alone.

Table 2 gives average bias, root mean squared error, and the Nash-Sutcliffe coefficient of efficiency for the estimations of mean field θ from the critical sampling points identified by $\mathbf{M}_T$ and $\mathbf{M}_E$. Estimating mean field θ from $\mathbf{M}_T$ resulted in the most inconsistency within the statistical indices. Estimations from one critical sampling point is supported by all indices over estimations from 2 and three critical sampling points identified with $\mathbf{M}_T$. When using four critical sampling points for estimation identified by $\mathbf{M}_T$, both estimations from SOM K-means and K-means improved significantly over estimations from a lower number of critical sampling points. The highest Nash-Sutcliffe value and lowest RMSE value for estimating mean field θ from any of the models is found when using the 4 critical sampling points identified from the SOM K-means algorithm using $\mathbf{M}_T$.

Estimations from critical sampling points identified by $\mathbf{M}_E$ show improved accuracy when increasing the number of points used in finding the mean field scale θ estimation. RMSE values decrease and NSCE values increase as more points are included. In comparison to the estimation accuracies from $\mathbf{M}_T$, improvement is seen except in estimation

from one point and in estimation from 4 points using the SOM K-means algorithm. Sampling only two points identified from $\mathbf{M}_E$ is more accurate than sampling three points using $\mathbf{M}_T$ from both methods and is more accurate than sampling four points using the K-means method for identifying critical sampling points.

Matrices $\mathbf{M}_T$ and $\mathbf{M}_E$ were analyzed separately in order to see the affects identification of sampling points, and thus field mean θ estimation, when including EMI data for each grid point. With the exception of the estimation from 4 critical sampling points identified by the SOM K-means algorithm from $\mathbf{M}_T$, the statistical indices in Table 2 support the use of identifying sampling points with $\mathbf{M}_E$. The relationship between soil texture values and the EMI values makes the value a strong factor in impacting θ values. The downfall with requiring this information is the added labor in finding the EMI values for points within the field. Ideally, only topographic data for θ estimation would be required. Topographic data can be quickly and efficiently measured at very high resolutions. Topographic data is readily available for much of the landscape in Iowa and in many other states. From the results in this research, in the absence of EMI data, more sensors will be needed at critical sampling points to accurately estimate mean field scale θ.

Comparison to the mean field scale estimate from the RSA OSLs shows the value of the new methods in estimating field scale θ from critical sampling points identified with topographic and EMI data. Using one point for estimation, the statistical indices support the use of the critical sampling points identified by the SOM K-means and K-means algorithms from $\mathbf{M}_T$ and $\mathbf{M}_E$ in all models. Using two points for estimation, sampling points identified using $\mathbf{M}_T$ were not supported for use over the RSA OSLs, but sampling points identified using $\mathbf{M}_E$ produced estimates better than using 4 OSLs to estimate mean field θ. Similarly,

using three points for estimation, sampling points identified by $\mathbf{M}_T$ were not supported for use over the RSA OSLs, but sampling points identified using $\mathbf{M}_E$ produced estimates better than using 4 OSLs. Finally, using 4 points for estimation, sampling points identified by $\mathbf{M}_T$ and $\mathbf{M}_E$ outperformed the RSA method in all statistical indices in all four scenarios. Given this information, the new methods for identifying critical sampling points based upon topographic and EMI data can be used to identify critical sampling locations to estimate the mean field scale θ with more accuracy than the RSA method of identifying OSLs. These results support the elimination of a dense sensor grid for mean field scale θ estimation that is required for finding OSLs from the RSA method.

Points identified for sampling

Completing the Rank Stability Analysis on the temporal θ data led to the selection of optimal sampling locations (OSLs) based on having the smallest deviation from the mean relative difference. Points 55, 23, 77 and 39 were the 4 points with the smallest standard deviation of the mean relative difference. All three points are in 138B Clarion loam soil. Points 23 and 55, the top two rank stable points, have the first and second highest elevation of all the points identified for sampling. Points 55 and 23 have the first and second lowest EMI values of any of the points identified for sampling with any of the algorithms (Table 4). Similarly, point 39, the fourth rank stable point, has the third highest elevation of all the points identified for sampling by all of the methods. The location of the points at the higher elevations is likely a factor in adding to the rank stability of the point over time. Point 77 is located near a transition between 138B and 55 soils and is the lowest in elevation (313.2 m) of all the points identified by the method.

Using $\mathbf{M}_\theta$, the SOM K-means algorithm and the K-means algorithm alone identified Point 77, the third highest rank stable location, as the one critical point to sample to estimate mean field scale θ. From Table 4, point 77 is one of the most commonly identified points by all of the algorithms. The algorithms selected the same 2 points (29 & 47) for sampling. With the exception of the value for slope aspect, points 29 and 47 are similar in all topographic and EMI index values. The algorithms identified the same 3 points for sampling (53, 67 & 83). Of these three points, point 53 has the highest EMI values and is the only point with a negative slope. Point 83 has a slope aspect value of ~40° meaning it receives less radiation from the sun than the other two points. In the identification of 4 critical sampling points, the only common point identified by both methods is point 65. Of all the points identified for sampling by all of the algorithms, point 65 has the highest H-H EMI and V-V EMI. High EMI values are associated with clay soils which have higher matric potential and thus more water holding capacity. The identification of this point for sampling allows points with higher clay content to be represented in calculating the mean field scale θ. The inclusion of this point likely adds to the increase in the NSCE for the model.

Recommended points for sampling identified from inputting $\mathbf{M}_T$ into the SOM K-means and the K-means algorithms are shown in Table 5. For one and two critical sampling point locations, the SOM K-means and K-means algorithms chose the same points for $\mathbf{M}_T$ (77; 51 & 67). In the identification of three critical sampling points from $\mathbf{M}_T$, points 21 and 67 are interchanged. From Table 4, points 21 and 67 have similar values for all variables except curvature. Points 21 and 41 are identified by both the SOM K-means and K-means methods when selecting 4 critical sampling point from $\mathbf{M}_T$. The methods differ in selecting

points 15 and 77 (SOM K-means) and points 51 and 59 (K-means). Point 15 has similar index values to those of 51 with the exception of slope aspect and EMI values.

In identifying one sampling location from $\mathbf{M}_E$, both of the algorithms choose point 21, a point that is similar to the average values of all the indices at the 42 grid points. In identifying two critical sampling locations from $\mathbf{M}_E$, the algorithms again identified the same locations (35 & 67). Point 35 has higher than average EMI values. In the identification of three critical sampling points from $\mathbf{M}_E$, the algorithms choose the same three points (51, 59, 67). These three points have the first, second, and fourth highest occurrences in being identified as critical sampling points for all the methods. Point 51 is the most common choice for sampling from $\mathbf{M}_T$ and $\mathbf{M}_E$. Of all the points identified for sampling from all of the methods, point 51 has the second highest EMI values. Similar to the selection of point 65 when the algorithms identified 4 critical sampling points from $\mathbf{M}_\theta$, the higher EMI values of this point correspond to a soil with higher clay content. Thus, point 51 likely represents points in the grid with that exhibit higher θ values throughout the measurement days.

The main conclusion in the above analysis is that sampling a higher number of points allows for the inclusion of sampling points with index values different from the average values of the indices at the grid points. Points identified using the RSA method are at higher elevations and have lower EMI values than the points identified with the SOM K-means and K-means algorithms with $\mathbf{M}_\theta$, $\mathbf{M}_T$ and $\mathbf{M}_E$ as inputs. Two of the points identified by the RSA method are never identified for sampling from any of the other algorithms. A complex statistical analysis of the points identified for sampling would be valuable in finding which factors are most influential in estimating θ values. This qualitative overview only hints at some of the discrepancies between the sampling points identified by the different algorithms.

One point worth considering when comparing the estimations based on time-series θ data to the estimations based on physical data is that the OSLs and BMUs from the θ data were found using only the 2004 season. This structure made the most sense chronologically for application of θ estimation. A variety of other variables not taken into account in this study could have been different in the 2004 season when compared to the other three years used for validation. Nothing from the precipitation data, sampling methods, or the time of year appears to be drastically different, but this point is worth considering when comparing the estimation of θ data to the estimations from physical data.

One of the reasons for the elimination of sampling days from the time-series data from each year was due to the sampling point being underwater. Eliminating these sampling days may have an impact on the pattern of θ behavior because the days with the highest average θ are not analyzed. Two days were eliminated from the 2004 temporal θ data, 4 days from 2005, 0 days from 2007, and 0 days from 2008. This would likely have the largest impact on the estimation methods based on the 2004 θ data rather than affecting the estimations from the algorithms based on the physical characteristics. Having data included in the 2004 θ data from the days when some of the points were under water may have an impact on the underestimates of the methods based on the time-series θ data.

The scale at which these estimations are made are not large based upon today's agricultural standards or compared to the resolution of some remote sensing devices (~18 acres). The size of the grid was chosen based upon the amount of time required for a person to collect data from all of the sampling points in less than two hours. This was done in an attempt to reduce the impact of drying during that time period. While discussing the size of the grid, one could question the resolution at which the samples are taken and speculate about

the inadequacy of the grid representing the actual θ of the entire field. The points cover a variety of different landscapes positions and soil types, but variability between the points could make the actual field θ average different than that found from the grid points.

The accuracy of the ThetaProbe could also be called into question when evaluating the legitimacy of these estimations. An R^2 value of 0.77 is high for a natural system, but also leaves room for error in the measurements. Human error is also a factor when sampling with the ThetaProbe. The location of the three samples taken at each grid point, the cleanliness of the probes of the instrument, the depth at which the instrument was placed in the ground all could have an effect on the accuracy of measurements.

After completing the θ estimations with both the SOM K-means algorithm and subsequently with the K-means algorithm alone, it appears from the resulting estimations that the classification into neurons on a SOM may not be needed to find critical sampling points based on temporal θ data or physical characteristics. With the exception of estimations from 3 and 4 points from $\mathbf{M}_T$, finding sampling points with the K-means algorithm results in estimation values with NSCE values equal to or greater than the values from the estimation from points identified for sampling from the SOM K-means algorithm. The strength of the SOM method is in organizing points on the output layer to produce a visual interpretation of the relationship between the points. The distance between the map neurons on the output layer gives insight into the θ behavior of sampling points in comparison to other points in the field. In addition, the number of clusters to be used for finding critical sampling points is supported by the u-matrices. Knowing that spatial relationships can be inferred from the output layer of SOMs, the SOM method is valuable in qualitatively understanding θ patterns

based on topographic and soil properties, but may not be needed to quantitatively identify sampling points.

Conclusion

The new methods proposed in this paper provide an effective way to estimate θ not only from past time-series θ behavior, but also from soil properties and physical characteristics. SOM K-means and K-means algorithms have the ability to identify critical sampling points using topographic and soil physical data that can be used to estimate mean field scale θ values. Being able to identify critical sampling points based solely upon physical characteristics that can be measured quickly and efficiently in comparison to *in-situ* θ measurements is a valuable outcome. The SOM algorithm, specifically the u-matrix output, is valuable in identifying the divisions within the input data. These divisions can then be used to divide the input data into similarly behaving clusters, and thus a representative point from those clusters can be identified for sampling. Results suggest that fewer critical sampling points are needed if EMI data is included in the field physical data for identifying critical sampling points as opposed to only using topographic data to identify sampling points. Moving forward, these results are promising in the pursuit to estimate mean field-scale θ without the need for extensive ground based θ sampling networks. Understanding and being able to accurately estimate θ is a key to understanding hydrologic performance in a wide range of natural modeling systems. Further studies are needed in order to validate these methods for finding critical sampling points in different environments and at different scales.

Tables and Figures

Table 1. Average bias, root mean squared error, and Nash-Sutcliffe efficiency index for mean field θ estimate from critical sampling points identified with $\mathbf{M}_\theta$

	# of Points	RSA $\mathbf{M}_\theta$	SOM $\mathbf{M}_\theta$	Kmeans $\mathbf{M}_\theta$
AB (cm^3/cm^3)	1	-0.010	-0.010	-0.010
	2	-0.016	0.003	0.003
	3	-0.014	-0.012	-0.012
	4	-0.013	0.005	-0.004
RMSE (cm^3/cm^3)	1	0.022	0.017	0.017
	2	0.021	0.014	0.014
	3	0.016	0.016	0.016
	4	0.016	0.010	0.012
NSCE	1	0.436	0.667	0.667
	2	0.483	0.754	0.754
	3	0.682	0.697	0.697
	4	0.708	0.876	0.832

Table 2. Average bias, root mean squared error, and Nash-Sutcliffe efficiency index for mean field θ estimate from critical sampling points identified with $\mathbf{M}_T$ (Topo) and $\mathbf{M}_E$ (Topo/EMI)

	# of Points	SOM $\mathbf{M}_T$	K-means $\mathbf{M}_T$	SOM $\mathbf{M}_E$	K-means $\mathbf{M}_E$
AB (cm^3/cm^3)	1	-0.010	-0.010	-0.004	-0.004
	2	0.018	0.018	0.002	0.002
	3	0.011	0.013	0.003	0.006
	4	-0.004	0.008	0.004	0.003
RMSE (cm^3/cm^3)	1	0.017	0.017	0.019	0.019
	2	0.025	0.025	0.013	0.013
	3	0.017	0.020	0.013	0.012
	4	0.008	0.013	0.011	0.010
NSCE	1	0.667	0.667	0.550	0.550
	2	0.264	0.264	0.798	0.798
	3	0.638	0.535	0.810	0.815
	4	0.923	0.785	0.850	0.891

Table 3. Points identified for sampling by the Rank stability analysis, SOM K-means, and K-means algorithms using $\mathbf{M}_\theta$ as input data

# of points	RSA $\mathbf{M}_\theta$	SOM $\mathbf{M}_\theta$	K-means $\mathbf{M}_\theta$
1	55	77	77
2	23,55	29,47	29,47
3	23,55,77	53,67,83	53,67,83
4	23,39,55,77	47,55,65,77	51,65,67,83

Table 4. Topographic and EMI data for any point identified for sampling by any of the methods. The point ID and the number of times the point was identified by the algorithms is in the first two columns.

BMU	Times Identified	Elevation (m)	Slope	Planar Curvature	Aspect (° CW from North)	H-H EMI (mS/m)	V-V EMI (mS/m)
1	3	313.7	1.6	0.1488	211.6	32.5	14.7
15	1	312.7	1.6	-0.0252	63.0	45.1	23.0
21	4	313.7	2.5	-0.0237	164.4	41.8	20.8
23	1	315.0	4.2	0.0267	126.5	27.8	12.2
29	2	313.1	2.5	0.0763	216.1	37.5	18.2
35	2	313.2	1.8	-0.0763	90.4	53.5	31.0
39	3	314.2	3.4	0.0755	124.1	33.1	15.0
41	3	314.4	2.0	0.1007	328.1	28.5	12.5
47	3	313.2	2.1	0.0740	320.2	43.1	23.9
51	9	312.6	1.8	-0.0237	153.3	69.3	43.6
53	2	313.4	1.8	-0.0755	166.0	52.6	29.1
55	2	314.6	2.3	0.0244	180.2	25.4	10.5
59	5	312.8	1.0	0.0511	124.4	42.2	20.8
65	2	312.6	1.7	-0.0999	216.5	77.8	48.2
67	10	313.7	2.9	0.0740	222.7	38.0	17.2
77	7	313.2	1.9	0.0259	221.3	33.6	16.2
83	3	313.5	3.2	0.0504	39.8	29.2	13.5
Mean all 42 grid points		313.3	2.1	0.0186	178.2	43.6	23.2

Table 5. Points identified for sampling by the SOM K-means and K-means algorithms using $\mathbf{M}_T$ (Topo) as input data and points identified for sampling by the SOM K-means and K-means algorithms using $\mathbf{M}_E$ (Topo/EMI) as input data

# of points	SOM $\mathbf{M}_T$	K-means $\mathbf{M}_T$	SOM $\mathbf{M}_E$	K-means $\mathbf{M}_E$
1	77	77	21	21
2	51,67	51,67	35,67	35,67
3	1,21,51	1,51,67	51,59,67	51,59,67
4	15,21,41,77	21,41,51,59	1,39,51,59	39,41,51,59

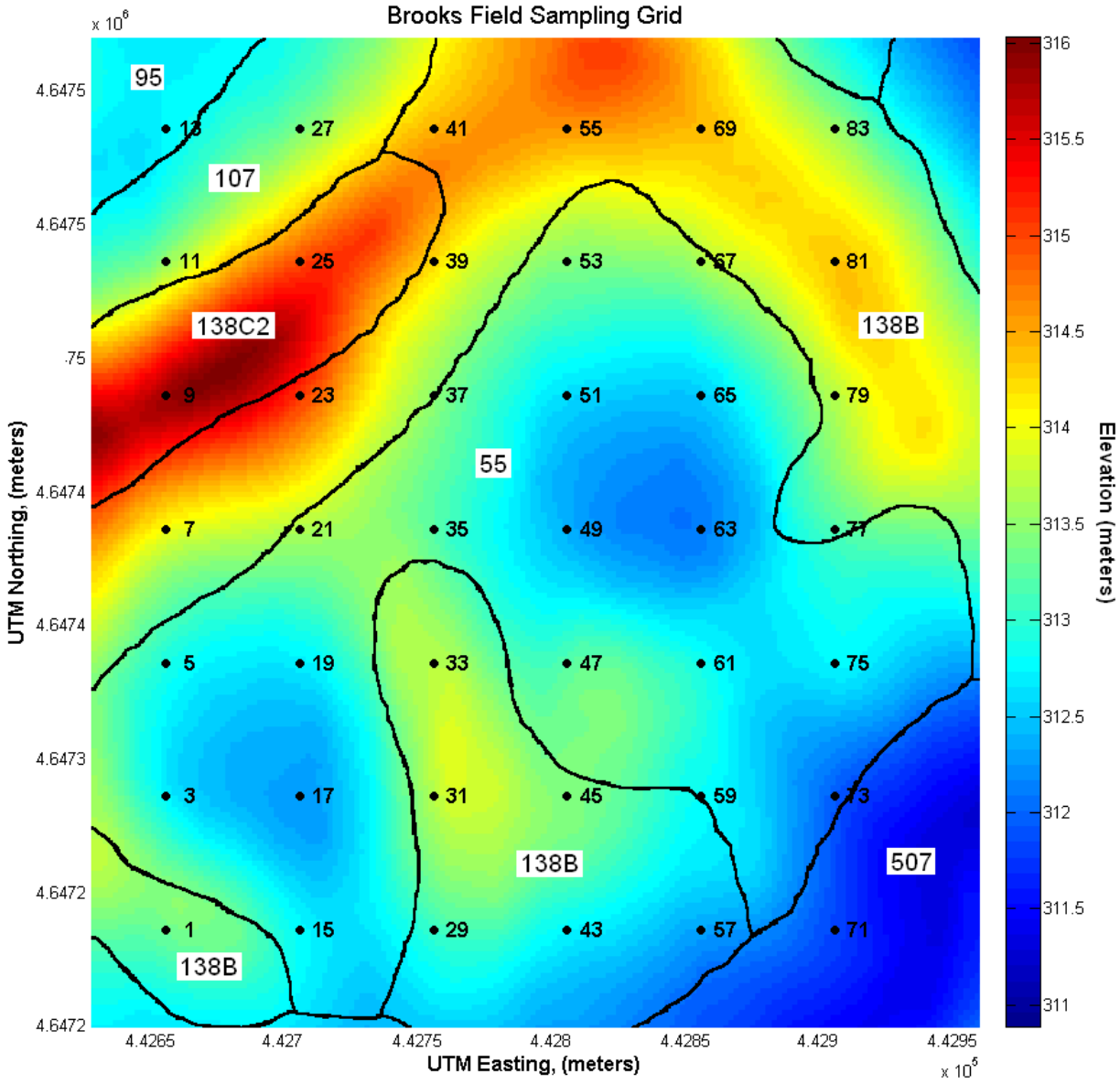


Figure 1. Brooks Field sampling grid with elevation and soil types. Points are on 50 meter spacing intervals. Soil type indices: 55: Nicollet loam, 1-3% slopes; 95: Harps loam, 1-3% slopes; 107: Webster clay loam, 0-2% slopes; 138B: Clarion loam, 2-5% slopes; 138C2: Clarion loam, 5-9% slopes, moderately eroded; 507: Canisteo clay loam, 0-2% slopes.

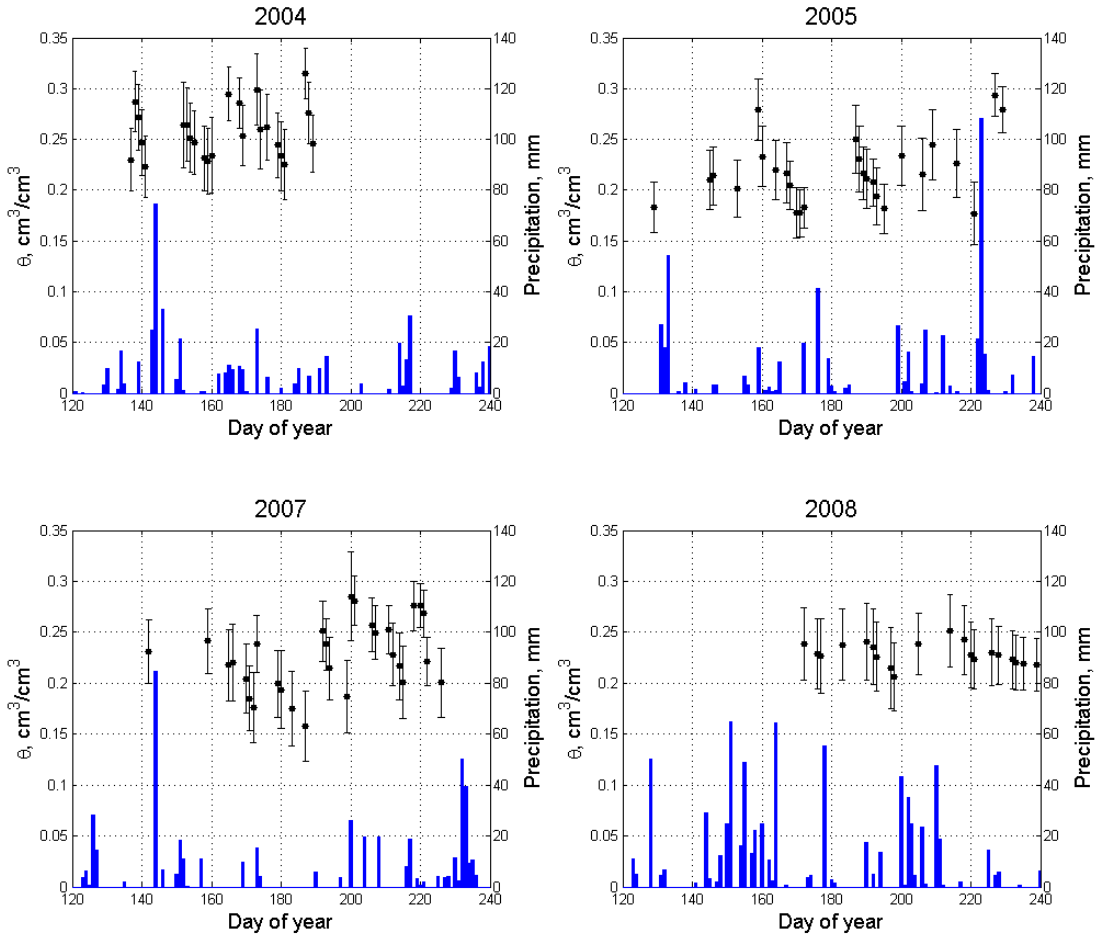


Figure 2. Average θ with standard deviation given by error bars for the 42 points within the Brooks field combined with precipitation data from the Ames 8 WSW stations at 42.0208 Lat, -93.7741 Lon.

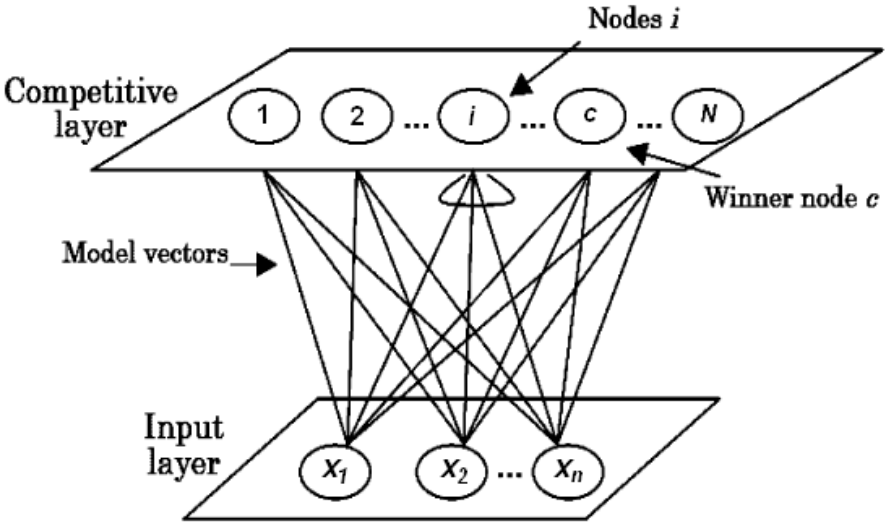


Figure 3. From Annas etal. (2007). Structure of a SOM with input layer (in this study temporal θ values and physical data) and competitive layer output layer, which produces an image similar to Figure 4.

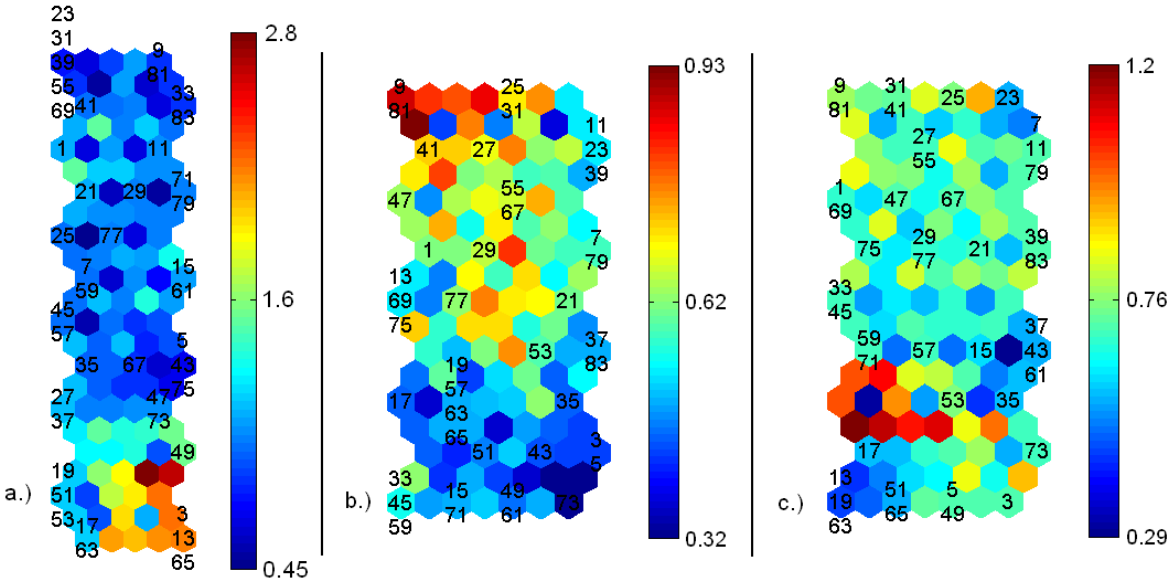


Figure 4. Unified distance matrices from inputting $\mathbf{M}_\theta$ (a), $\mathbf{M}_T$ (b), and $\mathbf{M}_E$ (c). Color bar to the right of each u-matrix denotes the Euclidean distance between neurons. Numbers correspond to point identification numbers.

References

Annas, S., T. Kanai, and S. Koyama, Principal Component Analysis and Self-Organizing Map for Visualizing and Classifying Fire Risks in Forest Regions, *Agricultural Information Research*, 16, 44-51, 2007.

Cereghino R., and Y. Park, Review of the Self-Organizing Map (SOM) approach in water Resources: Commentary, *Environmental Modeling and Software*, 24, 945-947, 2009.

Chang, D. H., Analysis and modeling of space-time organization of remotely soil moisture (Ph.D. dissertation), University of Cincinnati, 2001.

Famigletti, J., J. Rudnicki, and M. Rodell, Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas, *Journal of Hydrology*, 210, 259-281, 1998.

Honda, R. and O. Konishi, Temporal rule discovery for time-series satellite images and Integration with RDB, *Principles of Data Mining and Knowledge Discovery*, 2168, 2001.

Kaleita, A., J. Heitman, and S. Logsdon, Field Calibration for the Theta Probe for Des Moines Lobe Soils, *Applied Engineering in Agriculture*, 21(5), 865-870, 2005.

Kaleita, A., M. Hirschi, and L. Tian, Technical Note: Field Scale Surface Soil Moisture Patterns and Their Relationship to Topographic Indices, *Transactions of the ASABE*, 50(2), 2007.

Kalteh, A., Hjorth, P., and Berndtsson, R., Review of the self-organizing map (SOM) approach in water resources: Analysis, modeling and application, *Environmental Modeling & Software*, 23, 835-845, 2007.

Kohonen, T. *Self-Organizing Maps*. Springer, Verlag Berlin Heidelberg New York, 1995.

Lauzon, N., F. Anctil, and J. Petrinovic, Characterization of soil moisture conditions at temporal scales from a few days to annual, *Hyrdrological Processes*, 18, 3235-3254, 2004.

MacQueen, J., Some methods for classification and analysis of multivariate observation, *Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press 281-297, 1967.

McCuen, R. Z. Knight, and A. Cutter, Evaluation of the Nash-Sutcliffe Efficiency Index, Journal of Hydrology Engineering, 11, 2006.

Mele, M., and D. Crowley, Application of self-organizing maps for assessing soil biological Quality, *Agriculture, Ecosystems and Environment*, 126, 139-152, 2008.

Nash J., and J. Sutcliffe, River flow forecasting through conceptual models part 1-A discussion of principles, *Journal of Hydrology*, 10(3), 282-290, 1970.

Qiu, Y., B. Fu, J. Wang, and L. Chen, Soil moisture variation in relation to topography and land use in a hillslope catchment of the Loess Plateau, China , *Journal of Hydrology*, *240*, 243–263, 2001.

Robinson, D., C. Campbell, J. Hopmans, B. Hornbuckle, S. Jones, R. Knight, et al., Soil Moisture Measurement for Ecological and Hydrological Watershed-Scale Observatories: A Review, *Vadose Zone Journal,* 7(1), 358-389, 2008.

Romano, N. and M. Palladino, Prediction of soil water retention using soil physical data and terrain attributes, *Journal of Hyrdrology*, 265, 56-75, 2002.

Strobl, R., P. Robillard, R. Shannon, R. Day, and A. McDonnell, A Water Quality Monitoring Network Design Methodology for the Selection of Critical Sampling Points: Part 1, *Environmental Monitoring and Assessment*, 112, 137-158, 2006.

Vachaud, G., A. Passerat de Silans, P. Balabanis, and M. Vauclin, Temporal stability of Spatially measured soil water probability density function. *Soil Science Society of America Journal*, 49, 822-828, 1985

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, SOM Toolbox for Matlab 5, Technical Report A57, Neural Networks Research Centre, Helsinki University of Technology, Helsinki, Finland, 2000.

Western, A.W., R.B. Grayson, G. Bloschl, and G.R. Willgoose, T.A. McMahon, Observed spatial organization of soil moisture and its relation to terrain indices. *Water Resources Research, 35*(3), 797–810, 1999.

Yang, L., Spatio-temporal patterns of field-scale soil moisture and their implications for in situ soil moisture network design (Ph.D. dissertation), Iowa State University, 2010.

Yeh, P., and E. Eltahir, Stochastic analysis of the relationship between topography and the Spatial distribution of soil moisture, Water Resources Research, 34(5), 1251-1263, 1998.

CHAPTER 3: INVERSE DISTANCE WEIGHTING BASED UPON PHYSICAL CHARACTERISTICS FOR INTERPOLATION OF IN-FIELD SOIL MOISTURE

A paper to be submitted to *Journal of Hydrology*

Zach Van Arkel, Amy Kaleita, Brian Hornbuckle, Sally Logsdon

Abstract

The spatial and temporal variance of soil moisture complicates the ability to monitor and effectively predict soil moisture values. Identifying patterns and understanding the relationships between locations within a field is limited by the time and resources required for adequate monitoring. Remote sensing devices efficiently measure soil moisture over large areas, but the coarse resolution of measurements limits the use of the data. Finding a method to accurately estimate field scale soil moisture with limited in-field resources is the focus of this study. Given temporal soil moisture measurements at critical sampling locations throughout a field, the Euclidean distance can be found between these sampling points and all of the points in the field from their physical characteristics vector (elevation, slope, aspect, curvature, electromagnetic inductance) and used for interpolation of soil moisture values. Ultimately, this method can be used to find an accurate in-field soil moisture estimation without extensive monitoring.

Introduction

Soil moisture (θ) is of fundamental importance in crop and hydrology modeling and in weather prediction. The volume of θ is small in comparison to other water reserves in the hydrologic cycle, but plays an important role in these land-surface processes. Jaynes et al.

(2003) noted the response of crop yield to θ. Western et al. (1999) documented the ability to predict runoff at the catchment scale given θ values. Weather patterns are influenced by θ because surface soil water affects the energy exchange between the atmosphere and the land surface. The dependence of these models on values of θ at the required scale highlight the importance of accurate measurement of θ.

Unfortunately, θ measurements are often not available at the spatial resolutions adequate to capture the variability of the aforementioned processes. The low spatial resolution of remote sensing devices leads to an inability to capture different θ patterns at the field-scale. On the other end of the θ sensing spectrum, ground based sensors provide θ information at the point scale. Ground-based techniques are precise and can provide very high spatial resolution, but obtaining data at the field-scale with point scale values is time consuming and monetarily expensive because of the number of sensors required. Dense networks of sensors are needed to accurately capture spatial patterns of θ at the field scale (Li and Heap 2010). This lack of an efficient and high resolution method of measuring θ calls for the development of an interpolation method for estimation of θ patterns at the sub-field scale that does not require a dense sensor network.

Review of current interpolation methods

Similar to other environmental variables, the complexity and dynamic behavior of spatial θ patterns makes interpolation difficult. Of the interpolation techniques used for θ estimation, kriging is undoubtedly the most common. Li and Heap (2010) found that kriging methods outperform non-geostatistical interpolation techniques at various scales. Thattai and Islam (2000) used kriging methods to interpolate θ from remote sensing data in the Little

Washita watershed (~600 km^2). Bardossy and Lehman (1998) interpolated θ at the catchment scale using different variations of kriging. In a more applied study, Pandey and Pandey (2010) used kriging to predict θ for irrigation planning in a 2 hectare field from a 40 m x 40 m grid of *in situ* θ sensors.

Closely related to kriging, co-kriging has the ability to take into account secondary spatial information to aid in interpolation (Bishop and McBratney 2001). The different factors impacting θ can be included within the co-kriging algorithm to improve spatial estimtation. Yao et al. (2006) used co-kriging to interpolate θ by including micro-topography characteristics. Using co-kriging instead of kriging resulted in an improvement in the accuracy of estimating θ. Bardossy and Lehman (1998) also saw an improvement in interpolation quality when co-kriging using a topographic index as secondary information.

The influence of topographic features and soils data on θ calls for their inclusion in estimation of θ patterns. The topography of the landscape has an impact on flow channels, infiltration, potential radiation, and is related to the different soil types. Mohanty et al. (1997) found slope position to be the biggest contributor to temporal variability of θ. Not only does the slope affect the flow of the water during wet conditions, but slope position also has an impact on the potential radiation that can be received at each point. Western et al. (1999) found the best univariate predictor under wet conditions to be a function of the upslope contributing area. In the same study, the authors found potential radiation to be the best predictor of θ during dry conditions. Numerous other studies include different topographic characteristics in attempts to model and predict θ. (Yoo and Kim 2004; Western et al. 2001; Wilson et al. 2004, Famiglietti et al. 1998; Kim and Barros 2002; Mohanty and Skaggs 2001) Though each of these studies were completed on different spatial scales, all use the influence

of topographic features in θ estimation. Famiglietti et al. (1998) provides an in depth analysis of different topographic indices, how they are computed, and why they have an impact on θ patterns.

Differing soil types and textures will also have an impact on the spatial θ patterns. A worthy estimation of the soil type that can be found in one pass over the field is the electromagnetic inductance (EMI) (Tromp-van Meerveld and McDonnell 2009). The electrical conductivity correlates strongly with the soil particle size and texture (Grisso et al. 2009). The connection between soil texture and particle size with the hydraulic characteristics make EMI a valuable index in estimating θ. Khakural et al. (2008) found a linear relationship between electrical conductivity and soil water profile storage. The landscape and soil characteristics also correlated with the EMI measurements. Huth and Poulton (2007) found that EMI can provide quick and efficient means for monitoring θ in agroforestry systems.

A combination of factors that affect the θ is likely the answer to modeling the complex nature of θ patterns. Herbst et al. 2006 were best able to predict soil hydraulic properties at a point given the relative elevation, the slope, and the slope aspect. Although not attempting to predict θ, Green et al. (2007) used elevation, slope, aspect, curvature, and upslope contributing area in combination with spatial coordinates to predict crop yield. Mohanty and Skaggs (2001) noted the need to develop quantitative relationships between θ and various soil, topographic, and vegetation characteristics. Wilson et al. (2005) found a variety of terrain indices that had predictive power of θ patterns. In their concluding remarks, the authors state that spatial distribution of θ is not based on one terrain index but on a weighted combination of indices. Similarly, Western et al. (1999) describe an “index

approach" where a variety of different indices are found for points throughout the landscape and used for analysis. A combination of indices is needed to accurately estimate the dynamic behavior of θ.

Although co-kriging attempts to incorporate topographic and/or soils data, the current methods for interpolation of θ fall short in their need for dense sampling networks and their dependence on spatial relationships to accurately portray θ patterns. Using kriging methods, a variogram is first constructed from a dense network of observed data. Given the variogram, a spatial pattern is estimated from the data depending on how far an unknown point lays from a known sampling point (Western et al. 2001). When data variation is high and the data contains randomly distributed patterns, as can be the case with θ, the sampling density needs to be increased to capture the spatial changes within the landscape (Li and Heap 2010). Thus, the sampling density and spatial design will have an impact on the accuracy of the interpolation method. Abrupt changes within the landscape that have an impact on θ will not be sensed unless a dense sampling network is installed. Similarly, because spatial θ patterns are highly variable, points located near one another in the field may exhibit different θ values. The current methods rely on neighboring points to interpolate values at unknown points. This spatial dependence can lead to inaccuracies when sudden changes in θ exist due to changes in the landscape and soil characteristics.

Given the shortcomings of the current methods for interpolation of θ, a new method is needed to estimate θ values that bypasses spatial dependency and does not require a dense sensor network of θ values. The purpose of this research is to develop and test a new method for interpolation based only upon landscape characteristics that have an effect on spatial θ patterns. Using topographic and EMI data (as a proxy for soils information) for each point, in

the form of a vector for sampling points and unknown points, the Euclidean distance between the data vectors are used to interpolate values based upon an inverse distance weighting algorithm. This new vector space interpolation (VSI) method is valuable because it ignores spatial dependencies between estimated points and sampling points and because it eliminates the need for exhaustive pre-sampling with a dense sensor network.

Methods

Field data

This study analyzed three *in-situ* θ data sets with varying spatial scales at the Iowa Validation Site (IVS) in Story County near Ames, Iowa (Fig. 1). The site covers approximately 150 acres (~60 hectares) and contains soils common in the Des Moines lobe. Elevation at the site ranges from approximately 310 to 316 meters above sea level and is characteristic of the prairie pothole region in which it is located.

Topographic indices and EMI data make up the data vector that is used in the VSI method. The topographic indices used for analysis were found given the elevation data from the IVS collected at a ~20 m resolution with a GPS receiver on an all-terrain vehicle. The field was divided into 10 meter grid sections and the slope, planar curvature, and slope aspect (flow direction) for each grid point was derived from the elevation data using Surfer® (Golden Software, Inc., Golden, Colorado). The grid cell containing each of the sampling points was identified and the topographic indices for the sampling points were extracted from this information.

Electromagnetic inductance (EMI) data was gathered at a ~20 m resolution using an EMI sled pulled behind an all-terrain vehicle. Horizontal and perpendicular conductances in

units of milliSiemens/meter were found with this instrument and then interpolated at each 10 meter grid point and at each sampling point with the same methods described above. Thus, for each θ sampling point from the three data sets and each 10 meter grid point, values of elevation, planar curvature, slope aspect, horizontal EMI, and perpendicular EMI were available. These values composed the data vector that was used in the VSI method.

Three different sets of θ data with different time measurement and spatial locations are used for analysis. The sampling locations of each of the data sets can be seen in Fig. 1. The first set of data analyzed was gathered during the 2011 growing season using a neutron probe at 16 sampling locations throughout the field. Seven different measurement days are available and θ values from depth ranges of 0-10 cm and 10-20 cm are used in this study.

The second data set analyzed was also gathered during the 2011 growing season. Nine CS616 sensors were used to log θ data every hour from day of year 202 to 258. Values for θ from 11:00am, 12:00pm, and 1:00pm at each sensor were averaged and used for that day's θ value. In order to view the difference in θ estimation depending on the depth of the sensor, two different depths, 4.5 cm and 15 cm, are used in this study (Fig. 2). The locations of these sensors correspond with the location of the neutron probe sampling sites from the first data set (Fig. 1). The same UTM coordinates and topographic and EMI data vectors are used for analysis at corresponding neutron probe data sites.

The third data set was collected during the 2007, 2008, and 2010 growing seasons. An irregular time-series of θ measurements was gathered at 35 different sampling locations during this time span. In total 65 measurement days are used for analysis. The θ value used for analysis is an average of 3 samples taken within a ~0.5 m radius of each sampling location at a depth of 0-6 *cm* with a ThetaProbe moisture meter (Delta-T Devices, Cambridge

UK, marketed in the United States by Dynamax, Inc., Houston, Texas). Values from the probe were then converted to estimates of volumetric θ using a calibration developed for soils on the Des Moines lobe provided by Kaleita et al. (2005). A field calibration based on ThetaProbe measurements combined with gravimetric sampling resulted in a regression coefficient R^2 of 0.77.

From the 35 sampling stations described in data set 3, 3 points were identified as optimal sampling locations based upon the methods utilized in chapter 2. A method using self-organizing maps combined with a K-means, and a method using only a K-means clustering algorithm are used to find 3 optimal sampling locations based upon the topographic and EMI data of all 35 points. Three sampling points were identified based upon the classification of three soil types, clay, silt, and sand. This is consistent with other studies using self-organizing maps for classification into different textural groups (Chang 2001).

Each data set differs in sampling density, in the depth at which θ is measured and in the temporal variation of sampling. In addition, data set 3 builds on the methods presented in chapter 2 for identification of optimal sampling points for field-scale θ estimation. Identical techniques will be used to evaluate the estimations from data set 1 and data set 2. Because only 3 sampling sites are used to estimate θ values on the 10 m grid, the remaining 32 sampling points in data set 3 are used for validation of the interpolation method.

Vector space inverse distance weighting interpolation method (VSI)

Using the data vectors (topographic indices and EMI) the Euclidean distance was calculated between each sampling point and each 10 m grid point. The Euclidean distance is determined by:

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \tag{3.1}$$

where d denotes distance, $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors, and n is the dimension of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (n=6 in this application, dimensions are elevation, slope, planar curvature, slope aspect, horizontal conductance, and perpendicular conductance). For each θ data set, distances in the vector space were computed between each sampling point and the 10 meter grid points throughout the field. This distance was then used in an inverse distance weighting algorithm to find θ values at each point in the 10 meter grid. The formula for the inverse distance weighting algorithm is given below:

$$\theta(v) = \frac{\sum_{i=0}^{N} w_i(v)\theta_i}{\sum_{j=0}^{N} w_j(v)}, \text{where } w_i(x) = \frac{1}{d(v,v_i)} \tag{3.2}$$

where v denotes the interpolated point, v_i is a point with known θ (sampling locations), d is the Euclidean distance from observed point v_i to unobserved point v as calculated from equation (1), and N is the total number of observed points used in interpolation. Fig. 1 gives values for θ from VSI method for one day given θ values at the three critical points identified for sampling by the K-means algorithm.

Traditional IDW method

To compare the VSI method with the traditional inverse distance weighting method, equations (3.1) and (3.2) were again used but with geospatial distances instead. In equation (3.1), vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are 2-dimensional (the first dimension is UTM easting and the second UTM northing). Given this spatial distance, known θ values at the sampling locations could then be used to interpolate unknown θ values at the 10 meter grid points with equation (3.2).

Fig. 3 shows one day of interpolated θ values using the traditional IDW method given θ values at the three critical points identified for sampling by the K-means algorithm.

After introducing kriging in the introduction as the most common method for interpolation of θ values, the comparison of the VSI method to kriging would be expected. The vector method was not compared to kriging methods because the small number of sampling points in each data set eliminates the ability to construct a variogram. As the literature suggests, kriging methods will likely be more accurate than traditional IDW, but those methods require more dense sets of data for interpolation. The desire to use a small number of points to accurately estimate θ within the field does not lend support to kriging because a large number of sampling points in close spatial proximity are required.

Validation methods

To compare the accuracy of the VSI method with the traditional inverse distance weighting, leave one out (LOO) cross validation was employed for data sets 1 and 2. For data set 1, 15 sampling points were used to find θ values at the 16^{th} point; this was repeated 16 times so that each observed point was left out of the analysis once. Similarly, with data set 2, 8 sampling points were used to estimate the θ of the 9^{th} point. The estimations were then compared to the corresponding actual observed θ value. The Nash-Sutcliffe efficiency index and the root mean squared error are used to compare the different models (Nash and Sutcliffe 1970, McCuen etal. 2006). Values for these indices for the VSI method are given in Table 1 and for the traditional IDW in Table 2.

To validate estimates from data set 3, the 3 sampling locations used for interpolation were left out of the actual grid θ values. Thus, 3 points were used to estimate the θ values at

the remaining 32 different points in the field. The estimated values from each interpolation method were compared with the actual values providing an opportunity to validate each estimation model. Table 1 gives values for statistical comparison of the VSI method as compared to the traditional IDW algorithm. To further compare the methods, the Pearson correlation coefficient for actual θ vs. estimated values of θ was found for each method on each measurement day for data set 3. Inconclusive evidence that the VSI method improved θ estimation prompted comparison of the methods with the Pearson coefficient. A scatter plot was constructed to compare the Pearson coefficient values for each interpolation method given sampling locations from the SOM or K-means clustering algorithm (Fig. 4).

Results and Discussion

The linear relationship between the actual and estimated values of θ from the VSI method using the 9 CS616 sensors from data set 2 can be seen in Fig. 2. As depth increases, the Nash-Sutcliffe efficiency index increases suggesting that the model is more accurate in estimation at deeper depths. This finding is consistent with the idea that as depth increases, θ variability decreases. The effect of precipitation, overland flow, and drying process associated with the soil surface is diminished at deeper depths. A smaller range of θ values with less variability leads to better estimation by both interpolation methods.

Taking into account the soils information at specific sampling sites (Table 3) gives insight into θ estimates by the VSI method. Viewing Fig. 2, the relationship with the 1:1 line for different sampling sites can be combined with the soils information to understand the estimated values of the model at the different sites. The model appears to consistently

underestimate or overestimate different sampling locations. At a depth of 4.5 cm, the method consistently underestimates θ values for sampling point 705. At a depth of 15 cm, the model again underestimates the actual values for point 705. Point 714 is also underestimated by the model at the 15 cm depth. At 4.5 cm, values at sampling site 714 are overestimated at low actual values and underestimated at high actual values. From Fig. 1, sites 705 and 714 are located at lower elevations and in depressions in the landscape. Table 3 gives information about soil characteristics of each sampling point. Of the sampling sites, site 705 has the second highest clay content. Site 714 has a very low sand content and the highest silt and clay content of all the sampling sites. This information may explain the underestimation of the values of θ at each of these sites. The higher clay content at these sites results in a soil that has the potential to hold more water. Though the EMI was used as a proxy for soils data, this may not have adequately captured the difference in soil texture at these points, and thus caused the model to underestimate θ values.

Fig. 1 and Fig. 3 show the estimated θ values given θ data from one measurement day from the three critical sampling points identified by the K-means clustering algorithm. A significant difference in estimated θ values for the different points throughout the field can be observed. In Fig. 1, the spatial θ pattern closely follows the contour lines showing its dependency on topographic data for interpolation. Unlike in Fig. 3, the three sampling points have no apparent θ pattern surrounding them despite their influence on the interpolated values. Soil moisture values change depending on the landscape position because the VSI method bypasses the spatial dependency apparent in both traditional IDW and kriging algorithms. Values appear to be the lowest on side slopes and highest in the depression areas. Little variation in θ values is seen between the hill tops and the side slopes in the landscape.

The traditional inverse distance weighting method (Fig. 3) yields a spatial pattern that is strongly influenced by the location of the 3 observed data points used for interpolation. The 3 sampling points can be easily identified in Fig. 3 because of the spatial θ pattern that has been estimated for that measurement day. The dependence on the spatial relationship of the points for interpolation can be clearly seen by the resulting θ pattern. Consistent with Fig. 1, point 4 exhibits the highest θ value of the identified sampling points and points 36 and 46 exhibit similar, drier θ values. Points located an equal spatial distance from point 4 and point 36 exhibit estimated θ values in the intermediate range of the θ color scale. Rings of equal θ values surround the sampling points further showing the reliance on spatial relationships for interpolation by the traditional inverse distance weighting method.

As with any interpolation method, the estimated values of θ for unsampled points throughout the field will be limited to the range of values exhibited by the sampling points. The color bar at the bottom of both Fig. 1 and Fig. 3 shows the range of θ values within the field. The range for the VSI method is smaller than the range given by the traditional IDW method. The method for selection of sampling points used in chapter 2 aims at finding a sampling point that will exhibit θ values from 3 different classes of points with similar θ behavior. This improves the likelihood of eliminating the points in the landscape with the highest and lowest θ values and thus makes accurate estimations of θ at those locations unlikely.

Comparing statistical index values in Tables 1 and 2 lends support to the use of the VSI method for interpolation. The Nash-Sutcliffe model efficiency index is used to evaluate the accuracy of hydrologic models. Values for this index can range from 1 to –infinity. A value of 1 corresponds to a model that perfectly estimates the observed values. In each of the

different data sets and depths chosen for analysis, the Nash-Sutcliffe index improved when using the VSI method. RMSE values were smaller for the VSI method in all data sets with the exception of the TDR SOM sampling points used for estimation.

Although estimated values from the VSI make more sense qualitatively because of their lack of spatial dependence on sampling locations, little improvement is seen in the Nash-Sutcliffe efficiency index. Specifically, only a small increase in the NSCE is observed when using the VSI method in data set 3. This result does not lend support to the VSI for smaller scale interpolation, but does support the methods from chapter 2 because the 3 sampling points selected can accurately estimate field scale θ with either interpolation method. Finding the average θ of the three sampling locations and then using that value for the estimation at each of the 32 other points on that day leads to an estimation with a NSCE value of 0.69. Both the VSI method and the traditional IDW method have high Nash-Sutcliffe values because they are finding a complex average of the 3 sampling locations. In addition, the increased number of measurements leads to a lack of detail being exposed within the estimation data. Given actual θ values at each 10 meter point, the model efficiency spread would likely widen between the VSI method and the traditional IDW method. The irregular spatial θ patterns estimated by the traditional IDW method would be exposed in the model efficiency index given a denser sensor network for validation.

To further compare the estimation of the VSI method with the traditional IDW method using the 3 optimal sampling locations identified by the SOM K-means and K-means algorithm, the Pearson correlation was calculated for the actual vs. estimated θ values for each of the 65 measurement days in data set 3 (Fig. 4). Higher values of the Pearson correlation coefficients for daily values of θ exhibited when using the VSI method lend

support to the new algorithm. Viewing Fig. 4, the range of values for the Pearson correlation coefficient for the VSI method with the K-means BMUs as sampling locations is ~0 to 0.65 and -0.1 to 0.5 with the SOM BMUs as sampling locations. For the traditional IDW, the Pearson correlation coefficient ranges from ~ -0.3 to 0.35 with the K-means BMUs as sampling locations, and ~-0.35 to 0.35 with the SOM BMUs as sampling locations.

Overall, the majority of correlation coefficients for the VSI method are higher than the correlation coefficients for the traditional IDW method. Only 12 of the 65 values of the correlation coefficients for the traditional IDW using the SOM BMUs as sampling points are above zero. Using the sampling points identified by the K-means algorithm, a negative correlation between the Pearson coefficients for the VSI method and the Pearson coefficients for the traditional IDW can be observed. The lowest values for the correlation coefficient using the VSI method correspond to the highest values of the Pearson correlation coefficient for the traditional IDW. Because of the spatial nature of the traditional IDW, this result may imply that when the field exhibits homogeneous θ values over the sampling grid, a traditional IDW will estimate values more accurately than the VSI method. After a rain or during drought periods when little variation in θ is seen over the landscape, the controlling factors of θ will have a smaller impact on the spatial θ patterns. Instead, any random samples from the field will be sufficient in estimating the θ on that measurement day because of the homogeneous conditions. Because the estimations of the VSI model rely on topographic features and EMI data for interpolation of θ values, the accuracy of the estimations will decrease when the points with different physical characteristics exhibit the same θ values. Identifying the days when the VSI method was outperformed by the traditional IDW method

and then finding factors that may have had an influence on θ values will be beneficial in understanding the patterns exhibited.

Conclusion

A method for interpolating θ values at the sub-field scale based upon topographic and EMI data is presented and compared to traditional, geospatial interpolation methods. This method has no reliance on spatial relationships of the sampling points to unknown points and does not require a dense network of sensors to monitor θ at multiple locations. When applied to three different θ data sets the following conclusions can be made:

1. The new method requires no spatial relationship information between the known sampling point values and the unknown values. This spatial independence allows points with similar soil types and topographic characteristics that are not spatially near one another to be assigned similar θ values.
2. Accuracy for estimation of the model improves as depth is increased. This is likely due to the decreased variability in θ as depth increases.
3. The estimation of θ values from the newly proposed method is different depending on the site and its characteristics. At some sites, the model consistently underestimates or overestimates the actual θ value. Two sites that were underestimated had the first and second highest clay content of all the sampling sites. Finding a link between the factors impacting how the model estimates θ at each site could lead to a calibration for the model depending on those factors.

4. Although the accuracy of both methods is difficult to validate at the scale that can be estimated given topographic characteristics and spatial relationships, the proposed method outperforms the traditional inverse distance weighting algorithm. The accuracy of the new method in comparison to the traditional techniques will likely improve with increased density of known samples within the landscape for validation.
5. The need for only a few points to accurately find the field-scale average θ presented in chapter 2 is supported by this study. The average of the points identified for sampling can be used to estimate the field-scale θ value for a given measurement day.

Overall, the goal of decreasing the amount of sensors needed for accurately estimating θ values at the field scale is supported by this new method. A onetime gathering of elevation and EMI data can be used to identify points for sampling and then that same data can be used for interpolation of θ values using the VSI method. This independence from dense sensor networks saves time in the field and money required to buy, install, and maintain the networks. Given θ data from a small number of points within the field, the new method has potential to be valuable in crop and hydrology modeling, in remote sensing validation, and in weather prediction that is dependent on θ values at the sub-field scale.

Tables and Figures

Table 1. NSCE, RMSE for Vector Space Interpolation method

Sensor: Depth (cm)	Total Points	Points used for Prediction	Measurement days	Nash-Sutcliffe	RMSE
Neutron Probe: 0-10	16	15	7	0.27	0.05
Neutron Probe: 10-20	16	15	7	0.28	0.05
CS616: 4.5	9	8	57	0.38	0.06
CS616: 15	9	8	57	0.51	0.05
TDR Kmeans: 0-5	35	3	65	0.73	0.03
TDR SOM: 0-5	35	3	65	0.76	0.03

Table 2. NSCE, RMSE for Traditional IDW

Sensor: Depth (cm)	Nash-Sutcliffe	RMSE
Neutron Probe: 0-10	0.12	0.05
Neutron Probe: 10-20	0.11	0.05
CS616: 4.5	0.22	0.07
CS616: 15	0.44	0.05
TDR Kmeans: 0-5	0.68	0.04
TDR SOM: 0-5	0.62	0.04

Table 3. Sampling site %sand, %clay, %silt

Site	Elevation (m)	% Sand	% Silt	% Clay
701	313.01	33.5	34.3	32.2
702	312.24	27.5	37.9	34.7
703	310.22	38.4	32.5	29.2
704	311.01	39.6	30.3	30.1
705	311.56	31.7	32.1	36.2
706	314.34	45.7	28.4	25.9
707	312.69	59.4	20.6	20.0
708	315.88	52.8	22.3	25.0
709	315.49	36.0	34.7	29.2
710	312.62	40.1	28.2	31.7
711	312.34	36.0	33.2	30.7
712	311.1	24.5	39.6	35.9
713	312.07	47.4	24.9	27.7
714	311.33	12.6	46.9	40.5
715	310.19	32.6	36.0	31.3

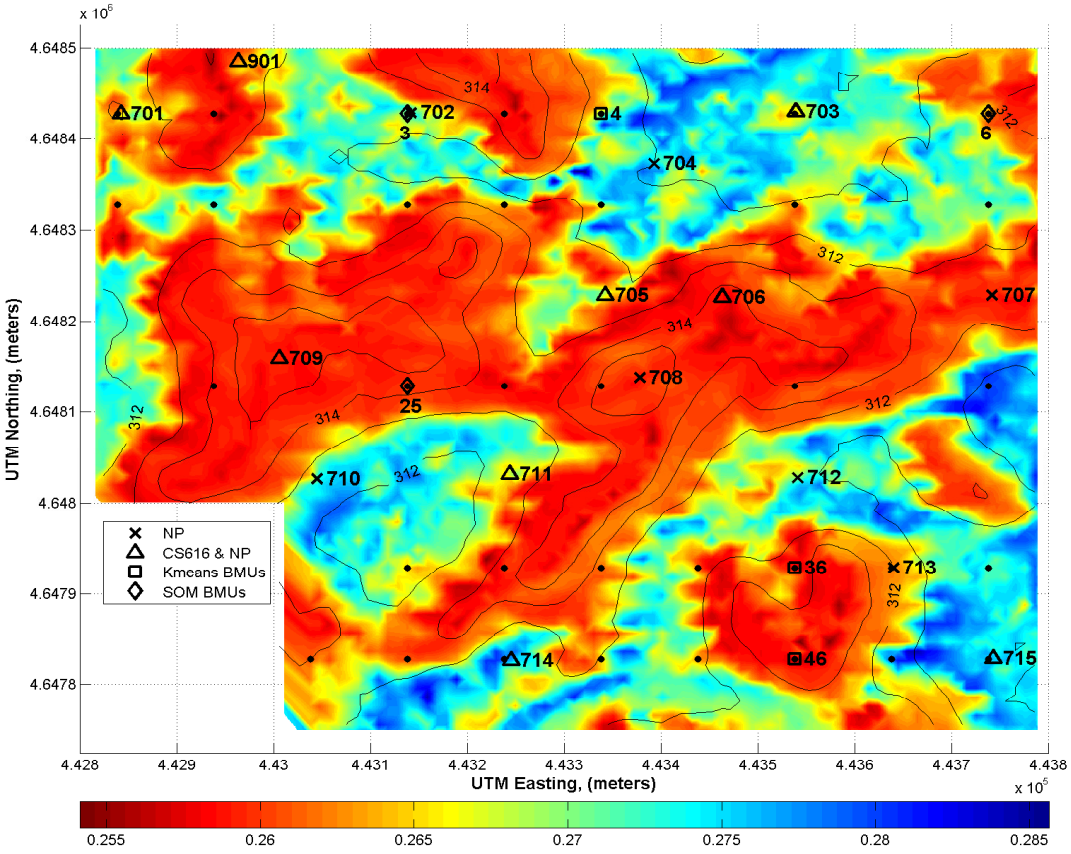


Figure 1. Vector space interpolation of θ values for one measurement day given θ values for that day from the 3 critical sampling points identified by the K-means clustering algorithm in chapter 2. Soil moisture values at 10 meter grid points are found using the VSI method and then linearly interpolated between 10 m grid points to create this map. Sampling locations of different sensors are given. Black circles correspond to the 35 point sampling grid from data set 3.

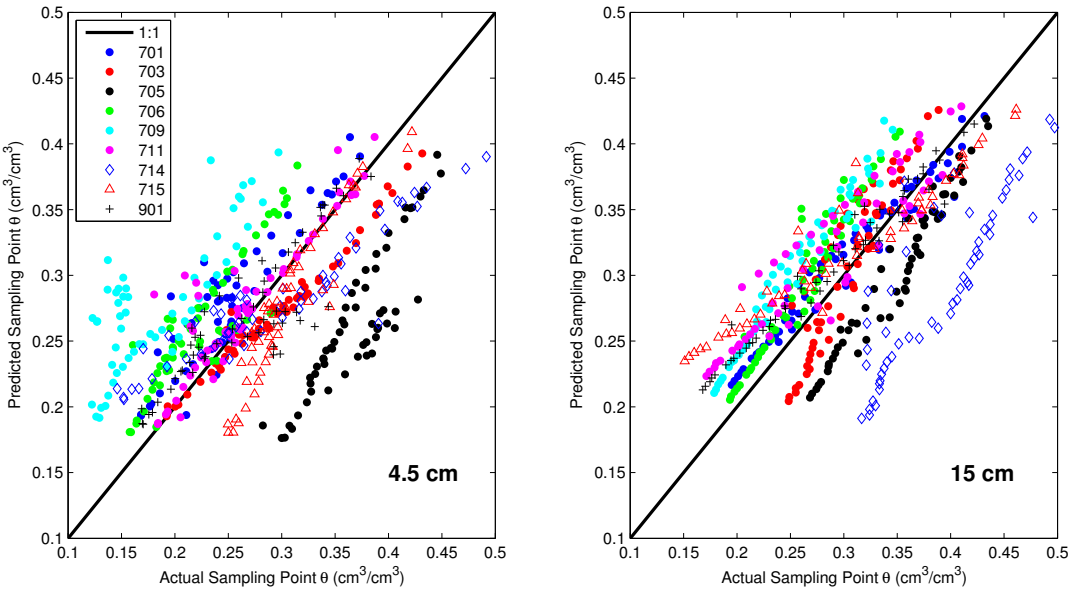


Figure 2. Actual vs. estimated θ value for different locations and depths of 9 CS616 sampling points at 4.5 cm and 15 cm depth

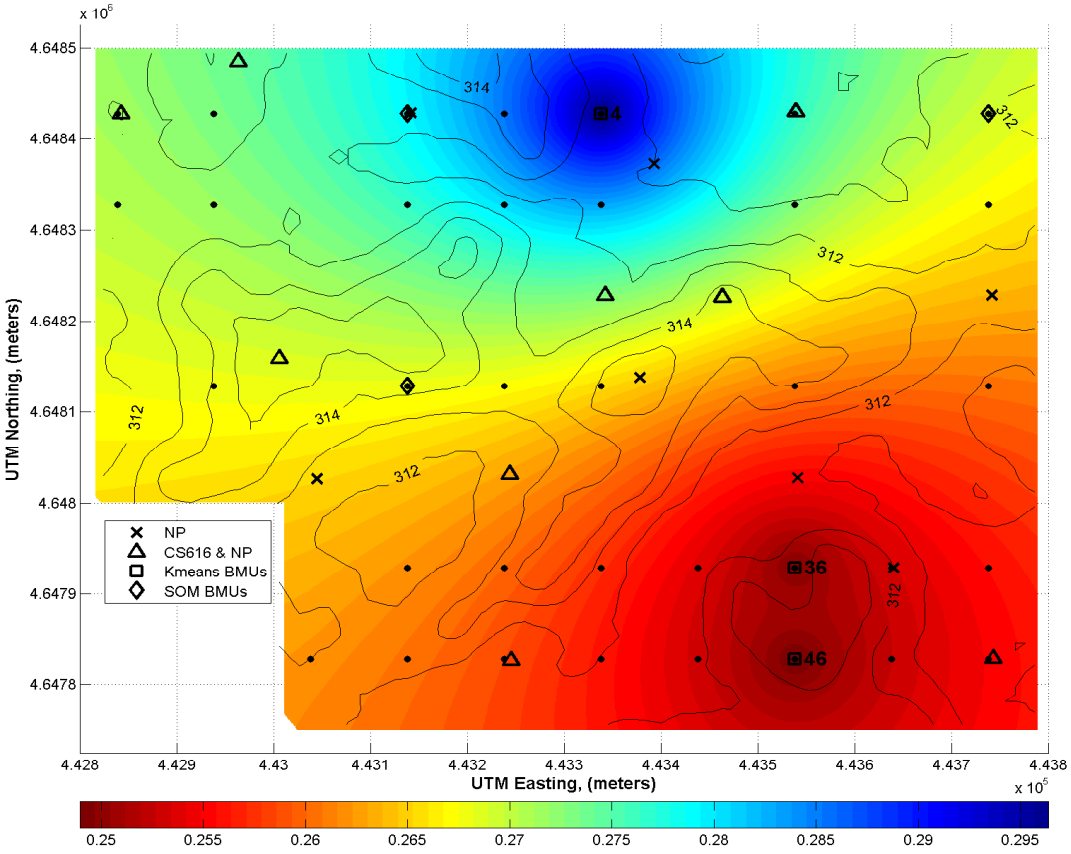


Figure 3. Traditional inverse distance weighting of θ values for one measurement day given θ values on that day from the 3 critical sampling points identified by the K-means clustering algorithm in chapter 2. Soil moisture values at 10 meter grid points are found using the traditional inverse distance weigthing algorithm and then linearly interpolated between the 10 m grid points to create this map. Sampling locations of different sensors are given. Black circles correspond to the 35 point sampling grid from data set 3.

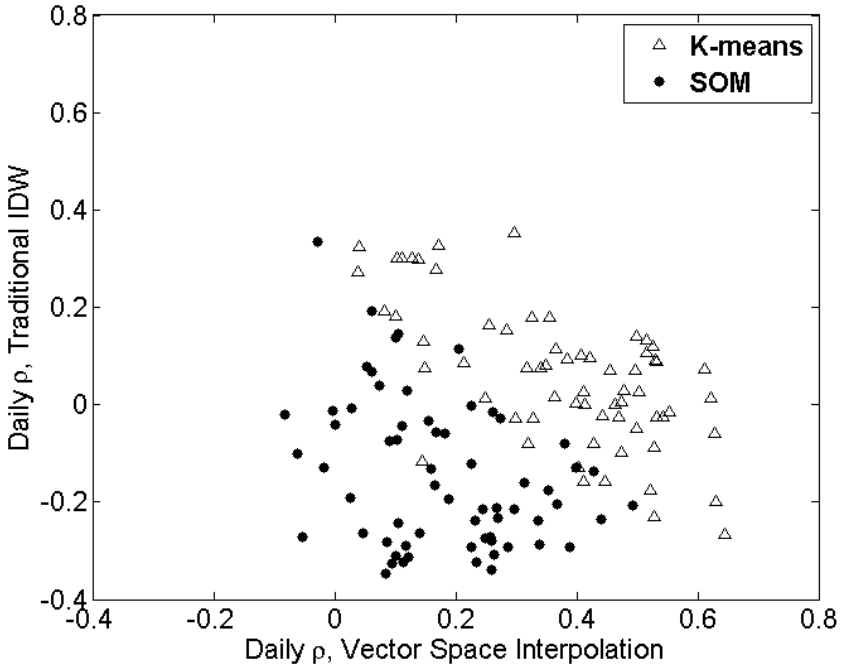


Figure 4. Pearson correlation coefficient values for vector space interpolation vs. Pearson correlation coefficient for traditional inverse distance weighting interpolation. Predictions using the sampling points identified by the K-means clustering algorithm and the SOM combined with the K-means clustering algorithm are used in finding the Pearson coefficient and separated in the figure.

References

Bardossy, A. and W. Lehman, Spatial distribution of soil moisture in a small catchment. Part 1: geostatistical analysis, *Journal of Hydrology*, 206, 1-15, 1998

Bishop T. F. and A. B. McBratney, A comparison of prediction methods for the creation of field-extent soil property maps, *Geoderma*, 103, 149-160, 2001.

Chang, D. –H., Analysis and modeling of space-time organization of remotely soil moisture (Ph.D. dissertation), University of Cincinnati, 2001.

Famiglietti, J., J. Rudnicki, and M. Rodell, Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas, *Journal of Hydrology*, 210, 259-281, 1998.

Grisso, R., M. Alley, D. Holshouser, and W. Thomason, Precision Farming Tools: Soil Electrical Conductivity, Virginia Cooperative Extension Publication 442-508, 2009.

Green, T. R., J. D. Salas, A. Martinez, and R. H. Erskine, Relating crop yield to topographic attributes using spatial analysis neural networks and regression, *Geoderma*, 2007.

Herbst M., B. Diekkruger, and H. Verrecken, Geostatistical co-regionalization of soil hydraulic properties in a micro-scale catchment using terrain attributes, *Geoderma*, 132, 206-221, 2006.

Huth, N. I., and P. L. Poulton, An electromagnetic induction method for monitoring variation in soil moisture in agroforestry systems, *Soil Research*, 45, 63-72, 2007.

Jaynes, D., T. Kaspar, T. Colvin, and D. James, Cluster analysis of spatiotemporal corn yield patterns in an Iowa field, *Agronomy Journal*, 95, 574-586, 2003.

Kaleita, A., J. Heitman, and S. Logsdon, Field Calibration for the Theta Probe for Des Moines Lobe Soils, *Applied Engineering in Agriculture*, 21(5), 865-870, 2005.

Khakural, B. R., P. C. Robert, and D. R. Hugins, Use of non-contacting electromagnetic inductive method for estimating soil moisture across a landscape, *Communications in Soil Science and Plant Analysis*, 29, 1998.

Kim, G. and A. Barros, Downscaling of remotely sensed soil moisture with a modified fractal interpolation method using contraction mapping and ancillary data, *Remote Sensing of Environment*, 83, 400-413, 2002.

Li, J. and A. Heap, A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors, *Ecological Informatics*, 6, 228-241, 2010

McCuen, R. Z. Knight, and A. Cutter, Evaluation of the Nash-Sutcliffe Efficiency Index, *Journal of Hydrology Engineering*, 11, 2006.

Mohanty, B. P., and T. H. Skaggs, Spatio-temporal evolution and time-stable characteristics of soil moisture within remote sensing footprints with varying soil, slope, and vegetation, *Advances in Water Resources*, 24, 1051-1067, 2001.

Mohanty, B. P., T. H. Skaggs, and J. S. Famiglietti, Analysis and mapping of field-scale soil moisture variability using high resolution, ground-based data during the Souther Great Plains 1997 (SGP97) Hydrology Experiment, *Water Resources Research*, 36, 1023-1031, 1997.

Nash J., and J. Sutcliffe, River flow forecasting through conceptual models part 1-A discussion of principles, *Journal of Hydrology*, 10(3), 282-290, 1970.

Pandey V. and P. K. Pandey, Spatial and Temporal Variability of Soil Moisture, *International Journal of Geosciences*, 1, 87-98, 2010.

Thattai, D. and S. Islam, Spatial analysis of remotely sensed soil moisture data, *Journal of Hydrologic Engineering*, October, 386-392, 2000.

Tromp-van Meerveld, H. J., and J. J. McDonnell, Assessment of multi-frequency electromagnetic induction for determining soil moisture patterns at the hillslope scale, *Journal of Hydrology*, 368, 56-67, 2009.

Western A., G. Bloschl and R. Grayson, Toward capturing hydrologically significant connectivity in spatial patterns, *Water Resources Research*, 37, 83-97, 2001.

Western, A.W., R.B. Grayson, G. Bloschl, and G.R. Willgoose, T.A. McMahon, Observed spatial organization of soil moisture and its relation to terrain indices. *Water Resources Research, 35*(3), 797–810, 1999.

Wilson D., A. Western, and R. Grayson, Identifying and quantifying sources of variability in termporal and spatial soi moisture observations, *Water Resources Research*, 40, 1-10, 2004.

Wilson D., A. Western, and R. Grayson, A terrain and data-based method for generating the spatial distribution of soil moisture, *Advances in Water Resources*, 28, 43-54, 2005.

Yao R., J. Yang, and L. Guang-ming, Spatial Variability of soil salinity and moisture and their estimations by co-kriging method – A case study in characteristic field of Yellow River Delta, *Journal of Soil and Water Conservation*, 2006.

Yoo, C. and S. Kim, EOF analysis of surface soil moisture field variability, *Advances in Water Resources*, 27, 831-842, 2004.

CHAPTER 4: GENERAL CONCLUSIONS

Conclusion

The goal of this research was to develop methods to efficiently and accurately estimate θ values at the field scale. Given the topographic and EMI data of the landscape where θ estimations are desired, a K-means clustering algorithm can be used to find optimal sampling locations. Whereas the rank stability analysis method requires temporal θ data from a dense grid of sensors to identify optimal sampling locations, the clustering algorithm only requires topographic and EMI data that can be obtained in one pass to identify sampling locations. A dense set of points with topographic and EMI indices within the landscape can then be divided into different clusters or families that will exhibit similar θ patterns. Finding these optimal sampling locations eliminates the need for dense sampling networks by allowing a small number of sampling points to find the same field-scale average as an entire sensor network. This data will be valuable for validation of remote sensing devices, as inputs in crop and hydrology models, and in weather prediction.

The importance of finding sampling points that adequately describe a cluster of points throughout the field is highlighted in this research. Using the K-means clustering algorithm, three different clusters were formed. The points with the topographic and EMI values that best matched the average of all other points in that cluster are chosen as critical sampling points in the landscape. Identifying three different points for sampling with this method helped realize the heterogeneity in θ that exists within a field. The complexity of spatial and temporal cannot be discovered if points with homogeneous landscape and soil types are used for sampling.

A second objective of this research was to develop a method to accurately estimate θ values at the sub-field scale given θ data from a small number of monitoring locations. A method was developed that depends on the topographic and EMI data for θ monitoring locations and the unknown points within the field to be interpolated. Because the method relied heavily on the topographic characteristics, the θ values were closely related to the topography of the field where the unknown values were interpolated. The scale at which the values of θ were interpolated is a higher resolution than the available θ monitoring grids, making validation of the method difficult. When divided into daily estimation of θ, the newly proposed vector space interpolation outperforms the traditional IDW method. The high variability in θ patterns currently requires dense sensor networks to adequately describe the differing θ values in the landscape. Linking θ values with the topography of the landscape leads to an independence from spatial relationships between known points for interpolation. This avoidance of spatial relationships allows differing topographic characteristics where changes in θ often occur to be the driving factor behind the interpolated value.

Prospects for future research

Although mentioned in the literature review, the mean moisture content was not used in the estimation methods in this research. The high values of the Pearson correlation coefficient for the traditional IDW interpolation corresponding with low values of the Pearson correlation coefficient for the VSI method promote the incorporation of mean moisture content into interpolation algorithms. Identifying the days when the Pearson correlation coefficient was low for the VSI method and finding a similarity between those days would be valuable in further understanding how mean soil moisture impacts field scale

θ values. Depending on the wetness conditions of the field of study, the estimation algorithm could adapt to the conditions. Spatial variability of θ values over the landscape resulting from a soaking rain or drought conditions will likely decrease making the use of topographic and soils data unnecessary. Similarly, different topographic characteristics or soils data may have a larger impact depending on the wetness conditions. Weights of different indices could be changed depending on the average θ values of the sampling locations. Future study on finding and incorporating this connection between the interpolation methods and wetness conditions is suggested.

In developing the vector space interpolation method the Euclidean distance was used because of its familiarity and simplicity. A variety of different formulas can be used to find the distance between two points in any dimension. The accuracy of estimating θ values using different distance formulas has the potential to increase. Maximizing the accuracy of these methods with different distance formulas is encouraged in future work.

Also installed at the IVS is a COSMOS probe that is used to remotely measure θ values with a footprint size of approximately 700 meters. A weighting function depending on the distance of interpolated points from the COSMOS sensor could be used to find an estimated COSMOS reading value from points interpolated using the vector space method. Similarly, a field scale soil moisture value could be estimated using the methods to find critical sampling points and a weighted average as used in the chapter 1. Given θ values at corresponding times from the COSMOS sensor and the installed θ sampling sites would allow for the validation of each method.

Values of elevation, curvature, slope, slope aspect, horizontal EMI, and perpendicular EMI were used in each application because the data was readily available. The topographic

indices can be quickly derived given the elevation data of the landscape. EMI requires another in field data gathering technique, but can be measured in one pass. Finding the indices that have the biggest impact on the estimation of θ values and eliminating the indices with little or no impact would save time and computing power in estimating θ values. Being able to estimate θ given only topographic information would be valuable in scenarios where measuring EMI data is not feasible (e.g. forested areas) Finding which indices have the biggest impact on the spatiotemporal θ estimation and making the model more efficient in terms of data required is another suggested path for future work.

It is important to remember that these methods were developed on fields where the elevation values varied by less than 10 meters. Using these models to estimate θ values when the topography is more variable is another suggested area of research. Sharper topographic features will likely have a more obvious impact on the estimating capability of the model. Specifically, the slope aspect will likely have a more pronounced impact on θ values when steeper slopes are present because of changes in potential evaporation. Outside the Des Moines lobe, topographic features are more variable and different cropping and hydraulic management strategies are used. Monitoring and testing these newly proposed methods in settings with different topographic features would further increase understanding of spatio-temporal θ patterns.

Further testing the accuracy of the vector space interpolation is another suggested area of research. Moran's I test was used on one day to find the patterns of error in estimation for the 32 points in data set 3. From this test it can be determined that the errors in estimation for the vector space interpolation method are randomly distributed throughout the field. Using the traditional IDW method the error values are closely clustered which is likely due to

the dependence of the method on the spatial relationship between sampling locations and points to be estimated. Appendix II gives an example hypothesis test with Moran's I method.

Future work could also be completed to determine at which scale three points can be used to interpolate θ values. Could the three points identified at the 150 acre IVS be used to estimate θ in the remaining 490 acres of the section in which it lays? For the eight surrounding sections? At a certain scale the differences in precipitation values will have a large impact on the variability of spatial θ patterns. Going even further, accurate precipitation data combined with optimal θ sampling points could be used to estimate θ values at large scales. Depending on precipitation an energy balance equation could be developed and used in θ estimation. The differing precipitation value would be just one more index that could be input in the θ estimation model. Finding the pixel size at which a small number of sampling sites could be used to interpolate values would be valuable in bridging the gap between θ sensing techniques and in models dependent on θ values.

APPENDIX I: MATLAB CODE

```
%Vector Space Inverse Distance Weighting

%% Load Been physical data
load('kentel elevation out.dat');
load('kentel EM hcon out.dat');
load('kentel EM pcon out.dat');
load('kentel plan curvature out.dat');
load('kentel terrain aspect out.dat');
load('kentel terrain slope out.dat');

%% Create 10 meter spacing grid of Been field

Been_Data_Grid =
horzcat(kentel_elevation_out(:,3),kentel_EM_hcon_out(:,3)...
    ,kentel_EM_pcon_out(:,3),kentel_plan_curvature_out(:,3),...
    kentel_terrain_aspect_out(:,3),kentel_terrain_slope_out(:,3),...
    kentel_terrain_slope_out(:,1),kentel_terrain_slope_out(:,2));

% eliminate field edges
[r,c]=find(Been_Data_Grid > 1e+37);
Been_Data_Grid(r(1:356),:)=[];

% cut out southwest corner
[r1,c1]= find(Been_Data_Grid(:,7)<443030 & Been_Data_Grid(:,8)<4648000);
Been_Data_Grid(r1(:,:),:)=[];

%%% Been_Data_Grid (the topo and EMI data for 10m points) is saved as a
.txt file (BeenPhysData 6.txt) so that it could be read in with
som_read_data and the normalize and pdist2 functions are be used. sD.data
and sP.data are used in calculating the distance in the vector space

%% load in Been physical data in 10 meter spacing
sD = som_read_data('BeenPhysData 6.txt',6);
%%% can change sD.data before making the SOM to change the parameters
%%% analyzed (elevation, EMI, etc)

% Normalize data but keep structure
sD = som_normalize(sD, 'var');

%% to find physical data values at sampling stations, this function finds
the value of the topo and EMI indices at the K-means BMUs UTM coordinates.
Been_35_SMGrid is a 35X2 matrix with easting and northing as columns
Kmeans_3BMUs =
vertcat(Been_35_SMGrid(16,:),Been_35_SMGrid(33,:),Been_35_SMGrid(3,:));
sp = Kmeans_3BMUs;

FElev =
TriScatteredInterp(Been_Data_Grid(:,7),Been_Data_Grid(:,8),Been_Data_Grid(
:,1),'nearest');
spElev = FElev(sp(:,2),sp(:,3));
```

```
FHEMI =
TriScatteredInterp(Been_Data_Grid(:,7),Been_Data_Grid(:,8),Been_Data_Grid(
:,2),'nearest');
spHEMI = FHEMI(sp(:,2),sp(:,3));

FPEMI =
TriScatteredInterp(Been_Data_Grid(:,7),Been_Data_Grid(:,8),Been_Data_Grid(
:,3),'nearest');
spPEMI = FPEMI(sp(:,2),sp(:,3));

FCurv =
TriScatteredInterp(Been_Data_Grid(:,7),Been_Data_Grid(:,8),Been_Data_Grid(
:,4),'nearest');
spCurv = FCurv(sp(:,2),sp(:,3));

FAspect =
TriScatteredInterp(Been_Data_Grid(:,7),Been_Data_Grid(:,8),Been_Data_Grid(
:,5),'nearest');
spAspect = FAspect(sp(:,2),sp(:,3));

FSlope =
TriScatteredInterp(Been_Data_Grid(:,7),Been_Data_Grid(:,8),Been_Data_Grid(
:,6),'nearest');
spSlope = FSlope(sp(:,2),sp(:,3));

sp_Data = horzcat(spElev,spHEMI,spPEMI,spCurv,spAspect,spSlope);

sP = som_data_struct(sp_Data); % create structure to use som_normalize
sP = som_normalize(sP, 'var');

%% Euclidean distance b/t 10m spacing points and Sampling Stations
dist = pdist2(sD.data,sP.data); % sD and sP structures not really
% needed, but used for convenience of using som_normalize function. The
.data appendix refers to the topo and EMI values at each point. sD.data is
7050 X 6 and sP.data is 3 X 6. dist is then 7050 X 3.

%% Soil Moisture Interpolation for one measurement
day = 24; %day can be chosen but must be >=3 because the first two columns
are easting and northing

surfaceSM = vertcat(Been_35_SM(5,:),Been_35_SM(33,:),Been_35_SM(23,:));  %
3X65 daily theta values of Kmeans BMUs
power = 1; % can change depending on IDW function

[r,c]=size(surfaceSM);
PointSM = [];

for k = 1:length(dist);
      SM = sum(((1./dist(k,1:r).^power).*surfaceSM(1:r,day)')/
      sum(1./dist(k,1:r).^power));
%%% inverse distance weighting to find theta at all points
```

```
      PointSM = vertcat(PointSM,SM);
%%% fill matrix with theta at each 10m point values
end



%% To find theta values for all 7050 points for all measurement days
surfaceSM = vertcat(Been_35_SM(5,:),Been_35_SM(33,:),Been_35_SM(23,:));
[r,c]=size(surfaceSM);
PointSM2 =[];  %%% create empty matrix for 10m theta values
for day = 3:c;
    PointSM = [];
        for k = 1:length(dist);
            SM = sum(((1./dist(k,1:r).^power).*surfaceSM(1:r,day)')/
            sum(1./dist(k,1:r).^power));
            %%% inverse distance weighting to find theta at all points
            PointSM = vertcat(PointSM,SM); %%% fill matrix with theta
values
        end
    PointSM2 = horzcat(PointSM2,PointSM);
end

ForSurferPlot = horzcat(Been_Data_Grid(:,7),Been_Data_Grid(:,8),PointSM2);
%%% ForSurferPlot is a 7050 by 67 matrix, easting and northing in the
first two columns and then 65 days of predicted θ values for all 10 meter
grid points
```

Reference

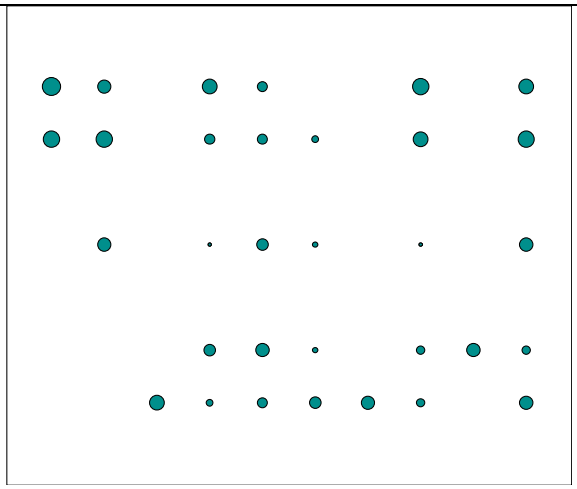
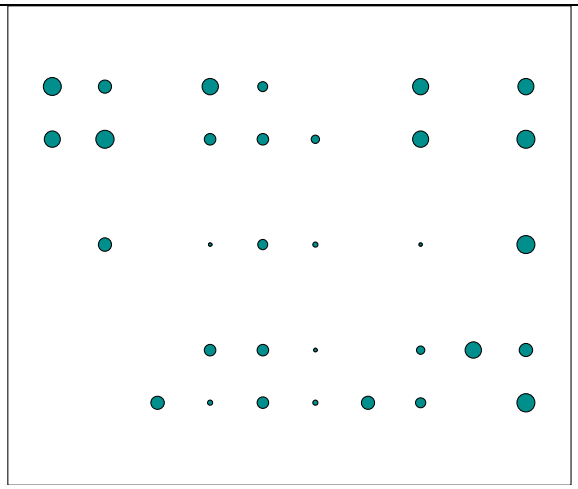
Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, SOM Toolbox for Matlab 5, Technical Report A57, Neural Networks Research Centre, Helsinki University of Technology, Helsinki, Finland, 2000.

APPENDIX II: MORAN'S I TEST

Evaluation of spatial distribution of the errors

- Graphical spatial distribution of the prediction errors

Below is the spatial distribution of the prediction errors obtained for each method from June 25 2007 after interpolation given values θ values at the K-means BMUs

Vector Space Interpolation	Traditional IDW
	
Small errors are spread over the whole area	Small errors are concentrated in the bottom part of the area

Interpretation:

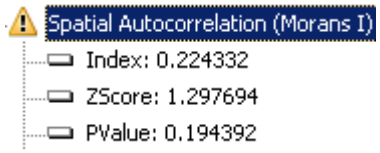
a. Big circles: high error values
b. Small circles: small error values

- Statistical test for the spatial distribution of the prediction errors: Moran's I test for autocorrelation

Hypotheses

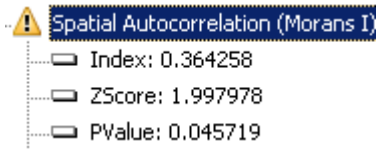
- H_0: there is no autocorrelation of prediction errors, that is, the prediction error values are randomly distributed in the whole area
- H_A: the autocorrelation is not equal to zero.

a. Vector space interpolation prediction errors



Conclusion: the p-value (0.19) for the z-score indicates that the autocorrelation of the prediction errors is zero, that is, they are randomly distributed over the whole area.

b. IDW prediction errors



Conclusion: the p-value (0.05) smaller than the significance level indicates that z-score for the spatial autocorrelation index is significant, that is, the autocorrelation is not equal to zero. This results implies that the prediction errors are not randomly distributed in the whole area.

Completed with assistance from Dr. Nerilson Terra Santos